Multinomijakauman ja Dirichlet-jakauman käytöstä bayesilaisessa päättelyssä

Pro gradu -tutkielma

Tiia Piipponen
Matematiikan ja tilastotieteen laitos
Helsingin yliopisto

28.3.2014

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Matematiikan ja tilastotieteen laitos

Tekijä — Författare — Author
Tiia Piipponen

Työn nimi — Arbetets titel — Title

Multinomijakauman ja Dirichlet-jakauman käytöstä bayesilaisessa päättelyssä

Oppiaine — Läroämne — Subject		
Matematiikka		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Maaliskuu 2014	35 s.

Tiivistelmä — Referat — Abstract

Tämän tutkielman tarkoituksena on antaa lukijalle pohjatiedot bayesilaisesta tilastollisesta päättelystä. Tutkielmassa esitellään kaksi usein bayesilaisessa päättelyssä käytettävää todennäköisyysjakaumaa: diskreetti multinomijakauma ja jatkuva Dirichlet-jakauma. Tutkielmassa tutustutaan jakaumien soveltamisen kannalta keskeisimpiin ominaisuuksiin. Lisäksi jakaumien yhteisjakauma Dirichlet-multinomijakauma esitellään.

Tilastollisen päättelyn tarkoituksena on tehdä havainnon perusteella johtopäätöksiä mittaamattomista suureista. Bayesilainen päättely sallii arvioitavaan suureeseen kohdistuvien ennakkotietojen huomioonottamisen päättelyprosessissa. Priorijakauma sisältää ennakkokäsitykset arvioitavasta suureesta. Uskottavuusfunktio on puolestaan havainnon todennäköisyysjakauma. Ennakkotieto ja havainnosta saatava tieto voidaan yhdistää Bayesin kaavalla ja näin priorijakauma päivitetään posteriorijakaumaksi. Johtopäätökset tehdään posteriorijakaumasta. Bayesilainen päättely vaatii paljon numeerista integrointia ja sen suosio on kasvanut tietoteknisen kehityksen myötä. Sitä sovelletaan nykyään monilla eri tieteenaloilla.

Tutkielmassa käsitellään sitä, kuinka multinomijakaumasta poimitun havainnon perusteella voidaan estimoida multinomijakauman parametreja bayesilaisin menetelmin. Tästä annetaan esimerkki. Dirichlet-multinomijakaumaa voidaan soveltaa esimerkiksi bayesilaisessa mallivertailussa ja myös tästä annetaan käytännön esimerkki.

Tutkielman sisällön ymmärtämiseksi vaaditaan, että lukijalla on pohjatiedot todennäköisyyslaskennan peruskäsitteistä. Lisäksi joukko-opilliset operaatiot ja tärkeimmät kuvauksiin, differentiaali- ja intergaalilaskentaan liittyvät asiat oletetaan tunnetuiksi. Tutkielmassa tehdään vertailua bayesilaisen ja frekventistisen tilastollisen päättelyn välillä ja oletetaan, että lukija tuntee frekventistisen tilastollisen päättelyn perusmenetelmät.

Avainsanat — Nyckelord — Keywords
Bayesilainen päättely, Multinomijakauma, Dirichlet-jakauma, Dirichlet-multinomijakauma

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

Luku 1

Johdanto

Tämän tutkielman tarkoituksena on antaa pohjatiedot bayesilaisesta tilastollisesta päättelystä. Lisäksi tutkielmassa tutustutaan kahteen usein bayesilaisessa päättelyssä käytettävään todennäköisyysjakaumaan: diskreettiin multinomijakaumaan ja jatkuvaan Dirichlet-jakaumaan. Tutkielman tarkoitus on, että lukija oppii jakaumien soveltamisen kannalta tärkeimmät ominaisuudet. Myös jakaumien yhteisjakauma Dirichlet-multinomijakauma esitellään.

Toisessa luvussa käydään läpi todennäköisyyslaskennan peruskäsitteitä. Lisäksi annetaan määritelmiä, jotka tuovat pohjatietoa myöhempien asioiden ymmärtämiselle ja joihin tullaan viittaamaan tutkielman muissa luvuissa. Kolmas luku käsittelee bayesilaisen tilastotieteen perusteita sekä tilastollista päätöksentekoa yleisesti. Neljännessä luvussa esitellään multinomijakauma ja sen käytöstä annetaan käytännön esimerkki. Lisäksi luvussa esitellään Dirichlet-jakauma ja sen ominaisuuksia. Multinomi- ja Dirichlet-jakaumien soveltamisesta bayesilaisessa tilastotieteessä annetaan esimerkki. Lopuksi luvussa johdetaan Dirichlet-multinomijakauman tiheysfunktio ja annetaan esimerkki sen soveltamisesta bayesilaisessa mallivertailussa.

Joukko-opilliset operaatiot ja niihin liittyvät laskusäännöt, sekä tärkeimmät kuvauksiin, sekä differentiaali- ja integraalilaskentaan liittyvät asiat oletetaan tunnetuiksi. Lisäksi keskeisimmät todennäköisyyslaskennan peruskäsitteet oletetaan tunnetuiksi, vaikka niitä hieman sivutaankin toisen luvun esitietojen yhteydessä. Lukijalle suositeltavaa pohjatietoa saa lähteistä [1] ja [2], joita on käytetty lähteenä tutkielman toisessa luvussa.

Tutkielmassa tehdään vertailua bayesilaisen ja frekventistisen tilastotieteen välillä ja oletetaan, että lukijalla on pohjatiedot frekventistisen tilastollisen päättelyn menetelmistä.

Luku 2

Esitietoja

2.1 Käsitteitä ja merkintöjä

Satunnaiskoe on idealisoitu koe, jolla on vähintään kaksi eri tulosmahdollisuutta. Kun kuvaillaan satunnaiskoetta, on aluksi määriteltävä *alkeistapaukset*, joilla tarkoitetaan kokeen kaikkia tulosmahdollisuuksia. Alkeistapausten yhdessä muodostamasta joukosta käytetään nimitystä *perusjoukko* ja sitä merkitään tutkielmassa symbolilla Ω.

Perusjoukon osajoukkoja $A \subset \Omega$ kutsutaan *tapahtumiksi*. Osajoukkojen kokoelma $\mathcal{F}$ on σ-algebra. Kun sanotaan, että tapahtuma $A \in \mathcal{F}$ sattuu, niin tarkoitetaan, että kokeen tulos ω kuuluu joukkoon A ja merkitään $\omega \in A$.

Satunnaismuuttuja X on kuvaus, joka liittää jokaiseen alkeistapaukseen $\omega \in \Omega$ täsmälleen yhden reaaliluvun $X(\Omega)$. Tutkielmassa satunnaismuuttujia merkitään yleensä isoilla kirjaimilla $(X, Y, Z, ...)$. Satunnaismuuttujien arvoja puolestaan merkitään pienillä kirjaimilla $(x, y, z, ...)$.

Jos $X_1, \ldots, X_k$ on k kappaleen jono satunnaismuuttujia, jotka ovat määriteltyjä samalla perusjoukolla Ω, niin vektori $\boldsymbol{X} = (X_1, \ldots, X_k)$ on *k-ulotteinen satunnaisvektori*. Tutkielmassa k-ulotteisesta satunnaisvektorista käytetään myös nimityksiä *k-ulotteinen satunnaismuuttuja*, *moniulotteinen satunnaismuuttuja* tai lyhyesti *satunnaisvektori*. Kun tutkielmassa mainitaan lyhyesti satunnaismuuttuja, tarkoitetaan 1-ulotteista satunnaismuuttujaa.

2.2 Diskreetit satunnaismuuttujat ja jakaumat

Jos satunnaismuuttujan arvojoukko $X(\Omega)$ on numeroituva tai numeroituvasti ääretön joukko $\{x_1, x_2, \ldots\}$ ja $\{X = x_k\} \in \mathcal{F}$ kaikilla k, kutsutaan satunnaismuuttujaa *diskreetiksi*. Satunnaisvektori on diskreetti, jos sen jokainen komponentti on diskreetti satunnais-

muuttuja.

Yksittäisen satunnaismuuttujan arvon todennäköisyyttä kutsutaan *pistetodennäköisyydeksi*. Olkoot satunnaismuuttujat $X_1, \ldots, X_k$ määriteltyjä diskreetissä perusjoukossa Ω. Satunnaisvektorin $\boldsymbol{X} = (X_1, \ldots, X_k)$ arvojoukko on joukko $\{(x_1, \ldots, x_k) \mid x_i \in X(\Omega)\}$ ja arvoa $(x_1, \ldots, x_k) \in \mathbb{R}^k$ vastaavaa pistetodennäköisyyttä merkitään

$$P\{X_1 = x_1, \ldots, X_k = x_k\} = p_{\boldsymbol{X}}(x_1, \ldots, x_k).$$

Pistetodennäköisyysfunktio määrittelee diskreetin k-ulotteisen satunnaisvektorin $\boldsymbol{X} = (X_1, \ldots, X_k)$ todennäköisyysjakauman. Satunnaisvektorin pistetodennäköisyysfunktion sijasta voidaan puhua myös satunnaismuuttujien $X_1, \ldots, X_k$ *yhteispistetodennäköisyysfunktiosta*.

Määritelmä 2.1. Olkoon $\boldsymbol{X} = (X_1, \ldots, X_k)$ diskreetti k-ulotteinen satunnaisvektori, jonka jokainen komponentti on määritelty perusjoukossa Ω. Merkitään satunnaisvektorin arvojoukkoa $\boldsymbol{X}(\Omega) = \{(x_1, \ldots, x_k) \mid x_i \in X(\Omega)\}$ kaikilla $i \in \{1, \ldots, k\}$. Kuvaus $f : \mathbb{R}^k \to \mathbb{R}$ on satunnaisvektorin $\boldsymbol{X}$ pistetodennäköisyysfunktio, jos

$$f(x_1, \ldots, x_k) = P\{X_1 = x_1, \ldots, X_k = x_k\} \quad \text{kaikilla} \quad (x_1, \ldots, x_k) \in \mathbb{R}^k \tag{2.2}$$

ja sillä on ominaisuudet

(i) $f(x_1, \ldots, x_k) \geq 0$ kaikilla $(x_1, \ldots, x_k) \in \mathbb{R}^k$,

(ii) $f(x_1, \ldots, x_k) = 0$, jos ja vain jos $(x_1, \ldots, x_k) \notin \boldsymbol{X}(\Omega)$,

(iii) $$\sum_{(x_1, \ldots, x_k) \in \boldsymbol{X}(\Omega)} f(x_1, \ldots, x_k) = 1.$$

Huomautus 2.3. Jos $\boldsymbol{X} = (X_1, \ldots, X_k)$ on diskreetti satunnaisvektori, niin todennäköisyys, että se saa arvon joukosta $A \subset \mathbb{R}^k$, voidaan laskea

$$P\{(X_1, \ldots, X_k) \in A\} = \sum_{(x_1, \ldots, x_k) \in A} f(x_1, \ldots, x_k) \quad \text{kaikilla } A \subset \mathbb{R}^k.$$

Määritelmä 2.4. Olkoon diskreetti k-ulotteinen satunnaisvektori $\boldsymbol{X} = (X_1, \ldots, X_k)$ kuten Määritelmässä 2.1. Satunnaisvektorin $\boldsymbol{X}$ kertymäfunktio on kuvaus $F : \mathbb{R}^k \to \mathbb{R}$, jolle

$$F(x_1, \ldots, x_k) = P\{X_1 \leq x_1, \ldots, X_k \leq x_k\} \quad \text{kaikilla } (x_1, \ldots, x_k) \in \mathbb{R}^k. \tag{2.5}$$

Jos satunnaiskokeeseen liittyvä satunnaismuuttuja voi saada täsmälleen kaksi eri arvoa, joiden todennäköisyydet pysyvät muuttumattomina toistettaessa satunnaiskoetta,

kutsutaan koetta *Bernoulli-kokeeksi*. Kokeeseen liittyvä perusjoukko voisi siten olla esimerkiksi {oikein, väärin}, {tapahtuu, ei tapahdu}, {kruuna, klaava}, {mies, nainen} tai {onnistuminen, epäonnistuminen}.

Määritelmä 2.6. Satunnaismuuttuja X noudattaa *Bernoulli-jakaumaa* parametrilla p $(0 \leq p \leq 1)$, merkitään $\mathrm{X} \sim Bernoulli(p)$, jos se voi saada kaksi erillistä arvoa, merkitään 0 ja 1, joiden todennäköisyysjakauma on

$$P\{X = 1\} = p \quad \text{ja} \quad P\{X = 0\} = 1 - p$$

ja sen pistetodennäköisyysfunktio on muotoa

$$f(x) = P\{X = x\} = p^x(1-p)^{1-x}, \quad \text{kun } x = 0, 1. \tag{2.7}$$

Binomijakauman pistetodennäköisyysfunktio määrittää satunnaismuuttujan arvoihin liittyvät todennäköisyydet, kun Bernoulli-koetta toistetaan n kertaa. Binomijakauma on erittäin tärkeä ja yleisesti käytetty jakauma tilastotieteessä. Sillä voidaan mallintaa riippumatonta toistokoetta, jossa yhdessä toistossa on aina tasan kaksi tulosmahdollisuutta.

Määritelmä 2.8. Satunnaismuuttuja X noudattaa binomijakaumaa parametrein n ja p $(n \in \{1, 2, \ldots\},\ 0 \leq p \leq 1)$, merkitään $X \sim Bin(n, p)$, jos sen pistetodennäköisyysfunktio on muotoa

$$f(x) = P\{X = x\} = \binom{n}{x} p^x(1-p)^{n-x} \quad \text{kaikilla } x \in \{0, 1, \ldots, n\}, \tag{2.9}$$

jossa $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ on *binomikerroin*.

2.3 Jatkuvat satunnaismuuttujat ja jakaumat

Satunnaismuuttuja on *jatkuva*, jos se voi saada mitä tahansa reaalilukuarvoja tietyltä väliltä ja satunnaisvektori on jatkuva, jos sen jokainen komponentti on jatkuva satunnaismuuttuja.

Tiheysfunktio määrittelee jatkuvan k-ulotteisen satunnaisvektorin $\boldsymbol{X} = (X_1, \ldots, X_k)$ todennäköisyysjakauman. Voidaan puhua myös satunnaismuuttujien $X_1, \ldots, X_k$ *yhteistiheysfunktiosta*.

Määritelmä 2.10. Kuvaus $f : \mathbb{R}^k \to \mathbb{R}$ on jonkin jatkuvan k-ulotteisen satunnaisvektorin $\boldsymbol{X} = (X_1, \ldots, X_k)$ tiheysfunktio, jos ja vain jos

(i) $f(x_1, \ldots, x_k) \geq 0$ kaikilla $(x_1, \ldots, x_k) \in \mathbb{R}^k$,

(ii) $f(x_1, \ldots, x_k)$ on integroituva yli $\mathbb{R}^k$:n,

(iii) $$\underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{k \text{ kpl}} f(x_1, \ldots, x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_k = 1.$$

Huomautus 2.11. Todennäköisyys P on mitta, jolla on ominaisuus

$$P\{(X_1, \ldots, X_k) \in A\} = \int_A \cdots \int f(x_1, \ldots, x_k) \mathrm{d}x_1 \cdots \mathrm{d}x_k \quad \text{kaikilla } A \subset \mathbb{R}^k$$

ja 1-ulotteisessa tapauksessa voidaan erityisesti laskea

$$P\{X \in [a, b]\} = \int_a^b f(x) \mathrm{d}x \quad \text{kaikilla } a, b \in \mathbb{R},\ a < b. \tag{2.12}$$

Määritelmä 2.13. Olkoon $\boldsymbol{X} = (X_1, \ldots, X_k)$ jatkuva k-ulotteinen satunnaisvektori, jolla on tiheysfunktio f. Satunnaisvektorin $\boldsymbol{X}$ *kertymäfunktio* on kuvaus $F : \mathbb{R}^k \to \mathbb{R}$, jolle

$$F(x_1, \ldots, x_k) = P\{X_1 \leqslant x_1, \ldots, X_k \leqslant x_k\} \quad \text{kaikilla } (x_1, \ldots, x_k) \in \mathbb{R}^k \tag{2.14}$$

ja tiheysfunktio f saadaan kertymäfunktiosta F kaavalla

$$f(x_1, \ldots, x_k) = \frac{\partial^k F(x_1, \ldots, x_k)}{\partial x_1 \cdots \partial x_k} \tag{2.15}$$

kaikissa pisteissä $(x_1, \ldots, x_k)$, joissa F:n osittaisderivaatat ovat olemassa. 1-ulotteisessa tapauksessa erityisesti

$$F(x) = P\{X \leqslant x\} = \int_{-\infty}^{x} f(t) \mathrm{d}t \quad \text{kaikilla } x \in \mathbb{R}. \tag{2.16}$$

Huomautus 2.17. 1-ulotteisen jatkuvan satunnaismuuttujan X tapauksessa kertymä- ja tiheysfunktiolle voidaan kirjoittaa myös

$$P\{X > x\} = 1 - P\{X \leqslant x\} = 1 - \int_{-\infty}^{x} f(t) \mathrm{d}t = 1 - F(x) \quad \text{kaikilla } x \in \mathbb{R}. \tag{2.18}$$

Beta-jakauma on jatkuva jakauma, jota tullaan tarvitsemaan myöhemmin. Sitä käytetään usein bayesilaisessa tilastotieteessä.

Määritelmä 2.19. Satunnaismuuttuja X noudattaa beta-jakaumaa parametrein α ja β ($\alpha > 0,\ \beta > 0$), merkitään $X \sim Beta(\alpha, \beta)$, jos sen tiheysfunktio on muotoa

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{kaikilla } 0 < x < 1, \tag{2.20}$$

jossa $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ on *beta-funktio* ja $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}\mathrm{d}x \quad (r > 0)$ on Eulerin *gamma-funktio*, jolle pätee $\Gamma(t) = (t-1)!$ kaikilla $t \in \mathbb{N}_+$.

Lause 2.21. *Beta-funktiolle pätee*

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x. \tag{2.22}$$

Todistus. Lause on suora seuraus Määritelmän 2.10 kohdassa (*iii*) annetusta tiheysfunktion ominaisuudesta, jonka nojalla

$$\int_{-\infty}^{\infty} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = 1.$$

Kertomalla puolittain beta-funktiolla saadaan $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = B(\alpha,\beta)$. □

2.4 Reunajakaumat ja satunnaismuuttujien riippumattomuus

Usein k-ulotteisen satunnaisvektorin $\boldsymbol{X} = (X_1, \ldots, X_k)$ jakauma tunnetaan, mutta halutaan keskittyä tarkastelemaan vain jonkin yksittäisen satunnaismuuttujan X_i jakaumaa. Tällaista jakaumaa kutsutaan satunnaismuuttujan X_i *reunajakaumaksi* ja se voidaan laskea satunnaisvektorin jakaumasta.

Määritelmä 2.23. Olkoon k-ulotteisen diskreetin satunnaisvektorin $\boldsymbol{X} = (X_1, \ldots, X_k)$ jokainen komponentti määritelty perusjoukossa Ω. Merkitään satunnaisvektorin arvojoukkoa $\boldsymbol{X}(\Omega) = \{(x_1, \ldots, x_k) \mid x_i \in X(\Omega)\}$ ja olkoon f satunnaisvektorin pistetodennäköisyysfunktio. Tällöin satunnaismuuttujan X_i reunajakauman pistetodennäköisyysfunktio f_i voidaan laskea kaavalla

$$\begin{aligned} f_i(x) = P\{X_i = x\} &= \sum_{(x_1,\ldots,x_k)\in\boldsymbol{X}(\Omega)|x_i=x} p_{\boldsymbol{X}}(x_1, \ldots, x_k) \\ &= \sum_{(x_1,\ldots,x_k)\in\boldsymbol{X}(\Omega)|x_i=x} f(x_1, \ldots, x_k). \end{aligned} \tag{2.24}$$

Diskreetissä tapauksessa satunnaismuuttujan arvon x_i reunatodennäköisyys saadaan siis laskemalla yhteen kaikkien sellaisten satunnaisvektorin arvojen $(x_1, \ldots, x_k)$ todennäköisyydet, joissa i:nnen satunnaismuuttujan arvo on x.

Jos satunnaisvektori on jatkuva, korvataan summaus integraalilla. Yksittäisen satunnaismuuttujan X_i reunajakauman tiheysfunktio saadaan siis jatkuvassa tapauksessa integroimalla satunnaisvektorin tiheysfunktiosta ne muuttujat, jotka eivät ole kiinnostuksen kohteena.

Määritelmä 2.25. Olkoon $\boldsymbol{X} = (X_1, \ldots, X_k)$ k-ulotteinen jatkuva satunnaisvektori ja olkoon f sen tiheysfunktio. Tällöin satunnaismuuttujan X_i reunajakauman tiheysfunktio f_i voidaan laskea kaavalla

$$f_i(x_i) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{k-1 \text{ kpl}} f(x_1, \ldots, x_k) \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \cdots \mathrm{d}x_k. \tag{2.26}$$

Jos $\{X_1, \ldots, X_k\}$ on joukko jatkuvia satunnaismuuttujia, joiden yhteisjakauma tunnetaan, voidaan laskea reunajakauma mille tahansa sen osajoukolle integroimalla yhteistiheysfunktio niiden muuttujien suhteen, jotka eivät ole kiinnostuksen kohteena.

Esimerkki 2.27. Olkoon $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)$ 5-ulotteinen satunnaisvektori, jolla on tiheysfunktio f. Satunnaismuuttujien X_1 ja X_4 2-ulotteinen reunatiheysfunktio voidaan laskea

$$f_{14}(x_1, x_4) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4, x_5) \mathrm{d}x_2 \mathrm{d}x_3 \mathrm{d}x_5.$$

Tutkielmassa tarkasteltavat satunnaismuuttujat ovat usein *riippumattomia*, joten satunnaismuuttujien riippumattomuus on vielä syytä määritellä.

Määritelmä 2.28. Olkoon $X_1, \ldots, X_k$ jono 1-ulotteisia diskreettejä (tai jatkuvia) satunnaismuuttujia, joiden yhteispistetodennäköisyysfunktio (tai yhteistiheysfunktio) on f. Olkoon kunkin satunnaismuuttujan X_i pistetodennäköisyysfunktio (tai tiheysfunktio) f_i, $i \in \{1, \ldots, k\}$. Satunnaismuuttujat $X_1, \ldots, X_k$ ovat riippumattomia, jos ja vain jos

$$f(x_1, \ldots, x_k) = f_1(x_1) \cdots f_k(x_k) \quad \text{kaikilla } (x_1, \ldots, x_k) \in \mathbb{R}^k. \tag{2.29}$$

Toisin sanottuna diskreetti (tai jatkuva) satunnaisvektori $\boldsymbol{X} = (X_1, \ldots, X_k)$ muodostuu riippumattomista komponenteista $X_1, \ldots, X_k$, jos ja vain jos satunnaisvektorin pistetodennäköisyysfunktio (tai tiheysfunktio) on komponenttien X_i reunajakaumien pistetodennäköisyysfunktioiden (tai tiheysfunktioiden) tulo.

Luku 3

Bayesilainen päättely

3.1 Tilastollisesta päättelystä

Kun tehdään tilastollista tutkimusta, on yleensä mahdotonta ja harvoin edes kannattavaa tutkia koko populaatiota, jolloin joudutaan tekemään johtopäätöksiä otoksen perusteella. Tilastollisen päättelyn tarkoituksena on tehdä mitatun tutkimusaineiston perusteella johtopäätöksiä sellaisista suureista, joita ei ole mitattu [3, s. 4]. Otoksesta mitattuja suureita kutsutaan *havainnoksi* ja havainnon perusteella voidaan tehdä päätelmiä populaation jakaumasta tai jakauman tuntemattomista parametreista. Päättelyä voidaan tehdä lukuisin eri menetelmin, joiden valinta riippuu siitä, mikä lähestymistapa käsitteen *todennäköisyys* tulkintaan valitaan.

Klassinen lähestymistapa perustuu yhtä todennäköisen periaatteeseen. Jos satunnaismuuttjalla X on n toisensa poissulkevaa yhtä todennäköistä tulosvaihtoehtoa $x_1, \ldots, x_n$, niin $P\{X = x_i\} = 1/n$, kun $i = 1, \ldots, n$. Lähestymistapaa on kritisoitu ensinnäkin siksi, että yhtä todennäköisen käsite liittyy todennäköisyyden käsitteeseen, jota pyritään määrittelemään, ja toiseksi siksi, että se ei tarjoa keinoa muiden kuin yhtä todennäköisiksi oletettujen tapahtumien todennäköisyyksien laskemiseen. [2, s. 3]

Frekventistisen lähestymistavan mukaan tapahtuman A todennäköisyys $P(A)$ määritellään tapahtuman sattumisen suhteellisena osuutena koetoistoista toistettaessa satunnaiskoetta äärettömän monta kertaa [2, s. 2]. Siis tehtäessä n riippumatonta koetoistoa, joissa tapahtuma A esiintyy n_A kertaa, niin $n_A/n \longrightarrow P(A)$, kun $n \longrightarrow \infty$. Frekventistinen määritelmä on kuitenkin liian epätäsmällinen, jotta sitä voitaisiin pitää tieteellisen todennäköisyysmääritelmän perustana, sillä sen mukaan todennäköisyys voidaan laskea vain toistettavissa oleviin kokeisiin liittyville tapahtumille [2, s. 3]. Sen avulla ei esimerkiksi voida arvioida todennäköisyyttä, että huomenna sataa.

Bayesilaisen lähestymistavan mukaan tapahtuman A todennäköisyys $P(A)$ edustaa todennäköisyyden laskijan omaa *uskomuksen astetta* tapahtuman sattumisesta [3, s. 11]

[4, s. 3]. Todennäköisyyden laskemisessa otetaan siis huomioon henkilön subjektiiviset ennakkokäsitykset tapahtumasta ja usein puhutaankin tietyn henkilön määrittelemästä subjektiivisesta todennäköisyydestä, ei tapahtuman *todellisesta* todennäköisyydestä [2, s. 4].

Keskeisin ero bayesilaisen ja frekventistisen päättelyn välillä on se, että ensimmäinen sallii arvioitavaan parametriin kohdistuvien ennakkokäsitysten ja -tietojen huomioonottamisen päättelyprosessissa, sillä parametri ajatellaan satunnaismuuttujana. Frekventistisen lähestymistavan mukaan ennakkokäsityksiä ei voida ottaa huomioon, sillä parametri itsessään ei ole satunnaismuuttuja, vaan kiinteä vakio, jonka todellinen arvo on tarkastelijalle tuntematon. [2, s. 328]

Bayesilaisessa tilastollisessa analyysissä tarvitaan paljon numeerista integrointia ja sen suosio on kasvanut tietoteknisen kehityksen myötä. Menetelmiä sovelletaankin nykyään yhä useammin esimerkiksi taloustieteessä, sosiaalitieteessä, lääketieteessä sekä koulutusta tai politiikkaa koskevissa tutkimuksissa. [3, s. 7]

3.2 Ehdollinen todennäköisyys ja Bayesin kaava

Tapahtuman A todennäköisyys $P(A)$ voidaan päivittää tapahtumalla B laskemalla todennäköisyys tapahtumalle A, kun tiedetään, että tapahtuma B on sattunut. Todennäköisyyttä kutsutaan tapahtuman A *ehdolliseksi todennäköisyydeksi* ehdolla B ja merkitään $P(A \mid B)$. [2, s. 49]

Määritelmä 3.1 (Ehdollinen todennäköisyys)**.** Olkoot A ja B satunnaiskokeen tapahtumia ja $P(B) > 0$. Tapahtuman A todennäköisyys ehdolla B on

$$P(A \mid B) = \frac{P(A, B)}{P(B)}, \tag{3.2}$$

jossa $P(A, B) = P(A \cap B)$ on tapahtumien A ja B yhteistodennäköisyys.

Ehdollinen todennäköisyys $P(A \mid B)$ voidaan ajatella *tapahtuman A todennäköisyytenä taustatiedon B valossa*. Taustatieto voi olla esimerkiksi aiempi tutkimustieto. Siten kaksi erilaisen taustatiedon omaavaa tarkastelijaa saattaa laskea eri todennäköisyyden samalle tapahtumalle [2, s. 4].

Uusien havaintojen avulla voidaan jatkuvasti pienentää tapahtumaan liittyvää epävarmuutta. Jos tapahtuman A todennäköisyys määritellään aluksi taustatiedosta B ehdollisena, merkitään $P(A \mid B)$, niin jälleen uuden tiedon C hankkimisen jälkeen voidaan tapahtuman todennäköisyys päivittää ehdolliseksi sekä taustatiedosta B, että uudesta tiedosta C, merkitään $P(A \mid B,C)$.

Jos tarkoituksena on estimoida tuntematonta parametria, voidaan uuden tiedon avulla oppia parametrista siten, että aiempi tieto päivitetään havainnosta saatavalla uudella tiedolla [4, s. 8]. Taustatieto ja uusi tieto voidaan yhdistää *Bayesin kaavalla* ja bayesilainen päättely perustuu juuri Bayesin kaavan käyttöön. Ennen Bayesin kaavaa todistetaan *kokonaistodennäköisyyden kaava.*

Lause 3.3 (Kokonaistodennäköisyyden kaava)**.** *Olkoon $B_1, \ldots, B_n$ satunnaiskokeen perusjoukon Ω ositus ja A epätyhjä perusjoukon Ω osajoukko. Siis kaikilla $i, j \in \{1, \ldots, n\}$, $i \neq j$, pätee $B_i \neq \emptyset$ ja $B_i \cap B_j = \emptyset$. Tällöin tapahtuman A todennäköisyydelle pätee*

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i). \tag{3.4}$$

Todistus.

$$\begin{aligned} P(A) &= P(B_1 \cap A) + \cdots + P(B_n \cap A) = P(B_1)P(A \mid B_1) + \cdots + P(B_n)P(A \mid B_n) \\ &= \sum_{i=1}^{n} P(B_i)P(A \mid B_i). \end{aligned}$$

□

Lause 3.5 (Bayesin kaava)**.** *Olkoot $B_1, \ldots, B_n$ ja A määritellyt kuten Lauseessa 3.3. Tällöin pätee Bayesin kaava*

$$P(B_j \mid A) = \frac{P(B_j)P(A \mid B_j)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)} \quad \textit{kaikilla } j \in \{1, \ldots, n\}. \tag{3.6}$$

Todistus. Yhtälön (3.4) mukaan $P(A) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i)$ ja yhtälöstä (3.2) saadaan $P(A \cap B_j) = P(A \mid B_j)P(B_j)$. Sijoittamalla nämä tulokset edelleen ehdollisen todennäköisyyden lausekkeeseen (3.2) saadaan

$$\begin{aligned} P(B_j \mid A) &= P(B_j \cap A)/P(A) = P(A \cap B_j)/P(A) \\ &= P(A \mid B_j)P(B_j)/\sum_{i=1}^{n} P(B_i)P(A \mid B_i). \end{aligned}$$

□

Esimerkki 3.7 (Esimerkki Bayesin kaavan käytöstä)**.** Oletetaan, että erään viruksen verikoetestin luotettavuus on 99%. Olkoon todennäköisyys, että satunnaisesti valittu henkilö kantaa virusta 1/20000. Mikä on todennäköisyys, että satunnaisesti valittu henkilö kantaa virusta, jos testitulos on positiivinen?

Ratkaisu:

Merkitään tapahtumia $B_1 =$ "henkilöllä ei ole virusta", $B_2 =$ "henkilöllä on virus" ja $A =$ "testi on positiivinen". Oletusten mukaan

$$P(B_1) = \frac{19999}{20000}, \quad P(B_2) = \frac{1}{20000},$$
$$P(A \mid B_1) = 0{,}01, \quad P(A \mid B_2) = 0{,}99.$$

Käytetään Bayesin kaavaa todennäköisyyden $P(B_2 \mid A)$ laskemiseksi:

$$\begin{aligned} P(B_2 \mid A) &= \frac{P(B_2)P(A \mid B_2)}{\sum_{i=1}^{2} P(B_i)P(A \mid B_i)} \\ &= \frac{P(B_2)P(A \mid B_2)}{P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2)} \\ &= \frac{\frac{1}{20000} \cdot 0{,}99}{\frac{19999}{20000} \cdot 0{,}01 + \frac{1}{20000} \cdot 0{,}99} \\ &\approx \underline{0{,}0049}. \end{aligned}$$

Siis todennäköisyys, että henkilöllä on virus, kun tiedetään, että testitulos on positiivinen, on vain noin 0,49%. Todennäköisyys saattaa vaikuttaa pienemmältä kuin olettaisi, sillä todellisuudessa tällaisia verikokeita ei yleensä tehdä satunnaisesti valituille henkilöille vaan ennemminkin sellaisille henkilöille, jotka uskovat, että heillä on riski kantaa virusta.

Esimerkin 3.7 todennäköisyyttä $P(B_2)$ kutsutaan *prioritodennäköisyydeksi*. Se on todennäköisyys tapahtumalle B_2 *ennen* tietoa tapahtuman A sattumisesta. Todennäköisyyttä $P(B_2 \mid A)$ kutsutaan puolestaan *posterioritodennäköisyydeksi*. Se on todennäköisyys tapahtumalle B_2 sen *jälkeen* kun tiedetään, että tapahtuma A on sattunut.

Bayesin kaavalla prioritodennäköisyydet $P(B_i)$ voidaan siis päivittää posterioritodennäköisyyksiksi $P(B_i \mid A)$, kun jokin tapahtuma A on havaittu. Bayesin kaavan prioritodennäköisyys $P(B_i)$ ilmaisee ennakkokäsityksen tapahtuman B_i todennäköisyydestä ennen tietoa tapahtuman A sattumisesta. Bayesilaisen päättelyn subjektiivisuus perustuu juuri siihen, että ennakkokäsitykset saattavat olla erilaisia ja bayesilaista lähestymistapaa onkin kritisoitu juuri subjektiivisuutensa vuoksi [2, s. 4].

3.3 Uskottavuusfunktio, priorijakauma ja posteriorijakauma

Bayesilaisessa päättelyssä tehdään johtopäätöksiä tuntemattomista suureista (parametri tai mittaamaton aineisto) ehdollistamalla ne havainnolla. Juuri havainnolla ehdollistaminen päättelyprosessissa erottaa bayesilaisen päättelyn muihin lähestymistapoihin perustuvista päättelymenetelmistä. [3, s. 7]

Jos X on diskreetti (tai jatkuva) satunnaismuuttuja, jolla on pistetodennäköisyysfunktio (tai tiheysfunktio) $g(x \mid \boldsymbol{\theta})$, jossa parametrivektori $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ on tuntematon ja jos $\boldsymbol{X} = (X_1, \ldots, X_n)$ on riippumaton satunnaisotos satunnaismuuttujan X jakaumasta, niin tällöin satunnaisvektorin $\boldsymbol{X} = (X_1, \ldots, X_n)$ pistetodennäköisyysfunktiota (tai tiheysfunktiota)

$$g(\boldsymbol{x} \mid \boldsymbol{\theta}) = g(x_1, \ldots, x_n \mid \boldsymbol{\theta}) = g(x_1 \mid \boldsymbol{\theta}) \cdots g(x_n \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta})$$

kutsutaan otoksen $\boldsymbol{X}$ *uskottavuusfunktioksi*. Uskottavuusfunktio on siis havainnon todennäköisyysjakauma, jossa tuntematon parametrivektori $\boldsymbol{\theta}$ käsitellään muuttujana ja havainto $\boldsymbol{x}$ on kiinteä. [2, s. 330-331]

Jakaumasta $g(x \mid \boldsymbol{\theta})$ poimitun otoksen perusteella voidaan arvioida, mitä arvoja tuntematon parametrivektori $\boldsymbol{\theta}$ todennäköisimmin saa. Estimoitava parametri on usein moniulotteinen ja jatkossa puhutaan myös moniulotteisen parametrin tapauksessa lyhyesti parametrista.

Ennen havainnon poimimista tilastotieteilijällä on usein aiempaan tietoon perustuvia ennakkokäsityksiä parametrin arvoista. Oletetaan, että nämä ennakkokäsitykset voidaan esittää parametrin $\boldsymbol{\theta}$ jakaumana $h(\boldsymbol{\theta})$. Tätä jakaumaa kutsutaan *priorijakaumaksi*, sillä se edustaa suhteellista uskottavuutta siitä, mikä on parametrin $\boldsymbol{\theta}$ todellinen arvo *ennen* havaintoa jakaumasta $g(x \mid \boldsymbol{\theta})$. [2, s. 327]

Jos tilastotieteilijällä ei ole ennakkotietoa parametrin käyttäytymisestä, voidaan priorijakauma muotoilla niin sanottuun epäinformatiiviseen muotoon. Tähän palataan tutkielman seuraavassa luvussa.

Otoksen poimimisen jälkeen prioritieto ja otoksesta saatu tieto voidaan yhdistää ja parametrin priorijakauma voidaan päivittää posteriorijakaumaksi. Lauseessa 3.5 esitetty Bayesin kaava kertoo, miten posteriorijakauman pistetodennäköisyysfunktio (tai tiheysfunktio) lasketaan havainnon poimimisen jälkeen [2, s. 330].

Lause 3.8 (Bayesin kaava jatkuville jakaumille)**.** *Olkoon $X_1, \ldots, X_n$ n kappaleen satunnaisotos jatkuvan satunnaismuuttujan X jakaumasta, jonka tiheysfunktio on $g(x \mid \boldsymbol{\theta})$. Otoksen uskottavuusfunktio on $g(\boldsymbol{x} \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta})$, jossa $\boldsymbol{x} = (x_1, \ldots, x_n)$ on havaintoarvot. Oletetaan, että k-ulotteinen parametri $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ on tuntematon ja jatkuva ja sen*

prioritiheysfunktio on $h(\boldsymbol{\theta})$. Posteriorijakauma parametrille $\boldsymbol{\theta}$ on

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})}{\int g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}. \tag{3.9}$$

Todistus. Koska otos on satunnaisotos satunnaismuuttujan X jakaumasta, niin $X_1, \ldots, X_n$ ovat riippumattomia ja samoin jakautuneita. Uskottavuusfunktio $g(\boldsymbol{x} \mid \boldsymbol{\theta})$ voidaan siten kirjoittaa muodossa

$$g(\boldsymbol{x} \mid \boldsymbol{\theta}) = g(x_1 \mid \boldsymbol{\theta}) \cdots g(x_n \mid \boldsymbol{\theta}).$$

Kertomalla uskottavuusfunktio ja prioritiheysfunktio keskenään saadaan $(n+k)$-ulotteinen yhteistiheysfunktio $g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})$. n-ulotteinen reunatiheysfunktio $m(\boldsymbol{x})$ havainnolle $\boldsymbol{x} = (x_1, \ldots, x_n)$ saadaan nyt integroimalla tulofunktio parametrin $\boldsymbol{\theta}$ yli

$$m(\boldsymbol{x}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{k \text{ kpl}} g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\theta_1 \cdots \mathrm{d}\theta_k. \tag{3.10}$$

Nyt yhtälö

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})}{\int g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \frac{g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})}{m(\boldsymbol{x})}$$

on Lauseessa 3.5 esitetyn Bayesin kaavan jakaumamuoto jatkuville jakaumille. □

Huomautus 3.11. Bayesin kaava voidaan samalla tavalla johtaa myös diskreettien satunnaismuuttujien ja parametrien jakaumille ja sellaisille jakaumille, joissa parametri on jatkuva ja satunnaismuuttuja, josta havainto poimitaan, on diskreetti. Jos satunnaismuuttuja X ja parametri $\boldsymbol{\theta}$ ovat diskreettejä ja otos $X_1, \ldots, X_n$ on poimittu X:n jakaumasta, niin posteriorijakauma saadaan kaavasta

$$f(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}} g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})}. \tag{3.12}$$

Yhtälössä (3.9) esitetyn Bayesin kaavan nimittäjää $m(\boldsymbol{x}) = \int g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ (tai $m(\boldsymbol{x}) = \sum_{\boldsymbol{\theta}} g(\boldsymbol{x} \mid \boldsymbol{\theta})h(\boldsymbol{\theta})$ diskreetissä tapauksessa) kutsutaan *reunauskottavuudeksi* tai *evidenssiksi* [4, s. 9]. Sen muoto riippuu siitä, onko parametrin $\boldsymbol{\theta}$ jakauma diskreetti vai jatkuva. Reunauskottavuus kertoo, kuinka todennäköistä on saada havainto $\boldsymbol{x}$, kun prioritieto parametrista $\boldsymbol{\theta}$ on annettu.

3.4 Bayesilainen todennäköisyysväli

Jos $[a, b]$ on luottamustason $(1-\alpha)$ frekventistinen luottamusväli 1-ulotteiselle parametrille θ, niin tarkoitetaan sitä, että poimittaessa useita otoksia, joista kustakin lasketaan tason

$(1-\alpha)$ luottamusväli parametrille, niin $100(1-\alpha)\%$ väleistä peittää parametrin *todellisen* arvon. Koska parametria ei käsitellä satunnaismuuttujana, ei ole oikein sanoa *"Parametri θ kuuluu välille $[a, b]$ todennäköisyydellä $(1-\alpha)$"*. Frekventistisessä luottamusvälissä välin päätepisteitä a ja b pidetään satunnaismuuttujina, ei parametria. [2, s. 412] [3, s. 112]

Kun bayesilaisessa päättelyssä parametri käsitellään muuttujana, jolla on todennäköisyysjakauma, voidaan laskea tason $(1-\alpha)$ väli $[a, b]$ siten, että pätee

$$P\{\theta \in [a, b] \mid \boldsymbol{x}\} = 1 - \alpha. \tag{3.13}$$

Bayesilaisessa päättelyssä ei puhuta luottamusvälistä, vaan voidaan käyttää esimerkiksi nimityksiä *bayesilainen todennäköisyysväli* tai *posterioriväli* [3, s. 4, 38]. Jos parametri on 1-ulotteinen, tehty havainto on $\boldsymbol{x}$ ja parametrin posteriorijakauma on $f(\theta \mid \boldsymbol{x})$, niin $100(1-\alpha)\%$ todennäköisyysväli $[a, b]$ saadaan yhtälöstä

$$P\{\theta \in [a, b] \mid \boldsymbol{x}\} = \int_a^b f(\theta \mid \boldsymbol{x}) \mathrm{d}\theta = 1 - \alpha. \tag{3.14}$$

Jos parametri on moniulotteinen, merkitään $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, voidaan yksittäisen θ_i:n todennäköisyysväli määrittää sille lasketusta reunaposteriorijakaumasta. Tästä tullaan antamaan esimerkki seuraavassa luvussa.

3.5 Mallien vertailu bayesilaisessa tilastotieteessä

Bayesilaisessa mallivertailussa on kyse siitä, että arvioidaan, kumpi kahdesta mallista kuvaa havaintoa paremmin ja käyttämällä Bayesin kaavaa voidaan laskea posterioritodennäköisyys kummalle tahansa malleista ehdolla havainto [5, s. 773]. Bayesilaisen mallivertailun frekventistinen vastine on hypoteesintestaus, jossa nollahypoteesi joko hylätään tai hyväksytään havainnon perusteella tietyllä merkitsevyystasolla [3, s. 250].

Olkoon $\boldsymbol{x} = (x_1, \ldots, x_n)$ n kappaleen otos satunnaismuuttujan X arvoja ja oletetaan kaksi vaihtoehtoista mallia M_1 ja M_2 havainnon $\boldsymbol{x}$ kuvailemiseen. Merkitään havainnon $\boldsymbol{x}$ todennäköisyystiheyttä mallin M_i mukaan $P(\boldsymbol{x} \mid M_i)$, $i \in \{1, 2\}$, ja merkitään mallien prioritodennäköisyyksiä $P(M_1)$ ja $P(M_2) = 1 - P(M_1)$. Nyt posterioritodennäköisyys mallille M_i ehdolla havaittu otos $\boldsymbol{x}$ voidaan laskea Lauseen 3.5 mukaisesti

$$P(M_i \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid M_i)P(M_i)}{P(\boldsymbol{x} \mid M_1)P(M_1) + P(\boldsymbol{x} \mid M_2)P(M_2)}, \quad \text{kun } i \in \{1, 2\}. \tag{3.15}$$

Mallien sopivuutta havainnon kuvaamiseen voidaan verrata toisiinsa laskemalla niin sanottu *Bayes-tekijä* jomman kumman mallin puolesta [5, s. 776].

Määritelmä 3.16 (Bayes-tekijä)**.** Olkoon $\boldsymbol{x} = (x_1, \ldots, x_n)$ n kappaleen otos diskreetin tai jatkuvan satunnaismuuttujan X arvoja ja merkitään otoksen $\boldsymbol{x}$ todennäköisyyttä mallin M_i mukaan $P(\boldsymbol{x} \mid M_i)$, kun $i \in \{1, 2\}$. Olkoot mallien prioritodennäköisyydet $P(M_1)$ ja $P(M_2) = 1 - P(M_1)$. Bayes-tekijä mallin M_2 puolesta voidaan laskea yhtälöstä

$$\frac{P(M_2 \mid \boldsymbol{x})}{P(M_1 \mid \boldsymbol{x})} = \text{Bayes-tekijä}(M_2; M_1) \times \frac{P(M_2)}{P(M_1)}. \tag{3.17}$$

Käyttämällä yhtälöä (3.15) saadaan posterioritodennäköisyyksien $P(M_2 \mid \boldsymbol{x})$ ja $P(M_1 \mid \boldsymbol{x})$ osamääräksi

$$\frac{P(M_2 \mid \boldsymbol{x})}{P(M_1 \mid \boldsymbol{x})} = \frac{P(\boldsymbol{x} \mid M_2)P(M_2)}{P(\boldsymbol{x} \mid M_1)P(M_1)}.$$

Sijoittamalla tämä yhtälöön (3.17) saadaan Bayes-tekijäksi mallin M_2 puolesta

$$\text{Bayes-tekijä}(M_2; M_1) = \frac{P(\boldsymbol{x} \mid M_2)}{P(\boldsymbol{x} \mid M_1)}. \tag{3.18}$$

Huomautus 3.19. Bayes-tekijä mallin M_2 puolesta voidaan laskea tiheysfunktioiden avulla reunauskottavuuksien osamääränä

$$\text{Bayes-tekijä}(M_2; M_1) = \frac{\int g(\boldsymbol{x} \mid \boldsymbol{\theta}_2, M_2) h(\boldsymbol{\theta}_2 \mid M_2) \mathrm{d}\boldsymbol{\theta}_2}{\int g(\boldsymbol{x} \mid \boldsymbol{\theta}_1, M_1) h(\boldsymbol{\theta}_1 \mid M_1) \mathrm{d}\boldsymbol{\theta}_1},$$

jossa $g(\boldsymbol{x} \mid \boldsymbol{\theta}_i, M_i)$ on mallin M_i mukainen uskottavuusfunktio otokselle $\boldsymbol{x}$ ja $h(\boldsymbol{\theta}_i \mid M_i)$ on mallin M_i mukainen priorijakauma parametrille $\boldsymbol{\theta}_i$, kun $i \in \{1, 2\}$. [5, s. 776]

Bayes-tekijällä verrataan siis toisiinsa saadun havainnon todennäköisyyksiä mallien M_2 ja M_1 mukaan laskettuna. Jos esimerkiksi mallin M_2 mukaan laskettu havainnon todennäköisyys on huomattavasti suurempi kuin mallin M_1 mukaan laskettu havainnon todennäköisyys (jolloin Bayes-tekijä$(M_2; M_1) >> 1$), niin on hyvin todennäköistä, että mallin M_2 uskottavuusfunktio kuvaa satunnaismuuttujan X jakautumista paremmin kuin mallin M_1 uskottavuusfunktio. On kuitenkin erilaisia määritelmiä siitä, mikä tässä tapauksessa on *huomattavasti suurempi todennäköisyys* ja *hyvin todennäköistä*. Tutkielmassa käytetään Taulukossa 3.5 esitettävää tulkintaa, jonka Robert Kass ja Adrian Raftery ovat määritelleet [5, s. 777]. Tulkinta Bayes-tekijän arvoista on esitetty taulukossa sekä englanniksi, että vapaasti suomennettuna.

Taulukko 3.5: Tulkinta Bayes-tekijän arvoista (Kass, Raftery, 1995)

Bayes-tekijä$(M_2; M_1)$	Evidence against M_1	Näyttö mallia M_1 vastaan
$1 - 3$	Not worth more than a bare mention	Ei juuri mainittava
$3 - 20$	Positive	Todellinen
$20 - 150$	Strong	Vahva
> 150	Very strong	Erittäin vahva

Luku 4

Multinomijakauma ja Dirichlet-jakauma bayesilaisessa päättelyssä

4.1 Multinomijakauma

Binomijakauma sopii vain sellaisten satunnaiskokeiden kuvailuun, joissa on tasan kaksi toisensa poissulkevaa tulosmahdollisuutta. Usein on kuitenkin tarpeen tarkastella sellaista satunnaiskoetta, jossa tulosmahdollisuuksia on enemmän kuin kaksi.

Oletetaan, että satunnaiskokeella on k toisensa poissulkevaa tulosmahdollisuutta, joita merkitään $E_1, \ldots, E_k$. Toistetaan koetta n kertaa ja määritellään satunnaisvektori $\boldsymbol{X} = (X_1, \ldots, X_k)$ siten, että X_i on tuloksen E_i esiintymisen lukumäärä toistokokeessa. Merkitään tuloksen E_i todennäköisyyttä $P(E_i) = p_i$ ja toistokokeesta saatua havaintoa $\boldsymbol{x} = (x_1, \ldots, x_k)$. Näin määriteltyä koetta kutsutaan *multinomikokeeksi*. Esimerkiksi nopanheitossa perusjoukko on $\Omega = \{1, 2, 3, 4, 5, 6\}$ ja jos määritellään 6-ulotteinen satunnaismuuttuja $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5, X_6)$ siten, että X_i edustaa silmäluvun i esiintymisen lukumäärää n:ssä heitossa, on kyseessä multinomikoe. Satunnaisesti valittujen ihmisten luokittelu silmien värin mukaan ja todennäköisyyksien laskeminen sille, että luokat ovat tietyn kokoisia, on käytännönläheinen esimerkki multinomikokeesta. *Multinomijakauma* on binomijakauman yleistys ja sillä voidaan mallintaa multinomikokeita.

Lause 4.1. *Olkoon satunnaiskokeella k toisensa poissulkevaa tulosmahdollisutta $E_1, \ldots, E_k$. Siis pätee $E_i \neq E_j$ kaikilla $i, j \in \{1, \ldots, k\}$, $i \neq j$. Merkitään tulosvaihtoehdon E_i sattumisen todennäköisyyttä $P(E_i) = p_i$ kaikilla $i \in \{1, \ldots, k\}$. Todennäköisyyksille p_i pätee*

$$p_1 + \cdots + p_k = \sum_{i=1}^{k} p_i = 1 \quad ja \quad p_i > 0 \quad kaikilla \quad i \in \{1, \ldots, k\}.$$

Toistetaan satunnaiskoetta n kertaa ja määritellään k-ulotteinen satunnaisvektori $\boldsymbol{X} = (X_1, \ldots, X_k)$ siten, että X_i on tulosvaihtoehdon E_i esiintymisen lukumäärä toistokokeessa kaikilla $i \in \{1, \ldots, k\}$. Koska tulosvaihtoehdot ovat toisensa poissulkevia, pätee

$$X_1 + \cdots + X_k = \sum_{i=1}^{k} X_i = n.$$

Näin määritellyllä satunnaisvektorilla $\boldsymbol{X}$ on jakauma, jonka pistetodennäköisyysfunktio on muotoa

$$\begin{aligned} f(x_1, \ldots, x_k) &= P\{X_1 = x_1, \ldots, X_k = x_k\} \\ &= \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i}, \end{aligned} \tag{4.2}$$

kun $x_1 + \cdots + x_k = n$. Sanotaan, että satunnaisvektori $\boldsymbol{X}$ noudattaa $(k-1)$-ulotteista multinomijakaumaa parametrein n ja $p_1, \ldots, p_k$, merkitään $\boldsymbol{X} \sim Mult(n; p_1, \ldots, p_k)$.

Todistus. (induktiolla)

Tapaus $k = 1$, $n > 0$ on triviaali.
Perusaskel.
Tapaus $k = 2$, $n > 0$:
Nyt pätee $X_1 + X_2 = n,\ p_1 + p_2 = 1$ ja

$$\begin{aligned} f(x_1, x_2) &= P\{X_1 = x_1, X_2 = x_2\} = P\{X_1 = x_1\} \cdot \underbrace{P\{X_2 = x_2 \mid X_1 = x_1\}}_{=1} \\ &= P(E_1 \text{ sattuu } n \text{ toistossa } x_1 \text{ kertaa}) \\ &= \frac{n!}{x_1!\,(n - x_1)!} p_1^{x_1} \left(1 - p_1\right)^{n - x_1}, \quad \text{kun} \quad x_1 \in \{0, 1, \ldots, n\}, \quad x_2 = n - x_1. \end{aligned}$$

Siis $\boldsymbol{X} = (X_1, X_2) \sim Mult(n; p_1, 1 - p_1)$ ja $X_i \sim Bin(n, p_i)$, kun $i \in \{1, 2\}$.
Induktioaskel.
Induktio-oletus:
Oletetaan, että Lause pätee, kun $k = K$, $n > 0$, eli
$\boldsymbol{X} = (X_1, \ldots, X_K) \sim Mult(n; p_1, \ldots, p_K)$ ja $f(x_1, \ldots, x_K) = \dfrac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}$, kun $x_1 + \cdots + x_k = n$.

Induktioväite:
Osoitetaan, että jos induktio-oletus pätee, niin Lause pätee, kun $k = K + 1$, $n > 0$:

$$\begin{aligned} f(x_1, \ldots, x_K, x_{K+1}) &= P\{X_1 = x_1, \ldots, X_K = x_K, X_{K+1} = x_{K+1}\} \\ &= P\{X_{K+1} = x_{K+1}\} \cdot P\{X_1 = x_1, \ldots, X_K = x_K \mid X_{K+1} = x_{K+1}\} \\ &= P(E_{K+1} \text{ sattuu } n \text{ toistossa } x_{K+1} \text{ kertaa}) \\ &\quad \cdot P(E_1 \text{ sattuu } n - x_{K+1} \text{ toistossa } x_1 \text{ kertaa}, \ldots \text{ ja } E_K \text{ sattuu} \\ &\quad\quad n - x_{K+1} \text{ toistossa } x_K \text{kertaa}) \\ &= \frac{n!}{x_{K+1}!\,(n - x_{K+1})!} p_{K+1}^{x_{K+1}} \cdot \frac{(n - x_{K+1})!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K} \\ &= \frac{n!}{x_1! \cdots x_K! x_{K+1}!} p_1^{x_1} \cdots p_K^{x_K} p_{K+1}^{x_{K+1}}, \end{aligned}$$

kun $x_1 + \cdots + x_k = n$. Koska induktioväite seuraa induktio-oletuksesta, on Lause tosi induktioperiaatteen nojalla.

□

Esimerkki 4.3 (Esimerkki multinomijakauman käytöstä)**.** Vuoden 2012 presidentinvaaleissa äänet jakautuivat ehdokkaiden kesken seuraavasti: Sauli Niinistö (N) 37,0 %, Pekka Haavisto (H) 18,8 %, Paavo Väyrynen (V) 17,5 %, Timo Soini (S) 9,4 %, Paavo Lipponen (L) 6,7 %, Paavo Arhinmäki (A) 5,5 %, Eva Biaudet (B) 2,7 % ja Sari Essayah (E) 2,5 % [6]. Mikä on todennäköisyys, että kolmesta satunnaisesti valitusta äänestäjästä kaksi äänesti Niinistöä, ja yksi Soinia?

Ratkaisu:

Tarkastellaan koetta kolmen toiston satunnaiskokeena, jossa jokaisessa toistossa valitaan satunnaisesti yksi äänestäjä.

Merkitään kokeen perusjoukkoa

$$\Omega = \{E_1 = \mathrm{N}, E_2 = \mathrm{H}, E_3 = \mathrm{V}, E_4 = \mathrm{S}, E_5 = \mathrm{L}, E_6 = \mathrm{A}, E_7 = \mathrm{B}, E_8 = \mathrm{E}\}$$

ja alkeistapausten sattumisen todennäköisyyksiä

$$p_1 = 0{,}370, \quad p_2 = 0{,}188, \quad p_3 = 0{,}175, \quad p_4 = 0{,}094,$$

$$p_5 = 0{,}067, \quad p_6 = 0{,}055, \quad p_7 = 0{,}027, \quad p_8 = 0{,}025.$$

Nyt satunnaisvektori $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$, jossa X_i on i:nnen alkeistapauksen esiintymisen lukumäärä kolmen toiston kokeessa, noudattaa multinomijakaumaa

parametrein $n = 3$ ja $p_1 = 0{,}370; p_2 = 0{,}188; p_3 = 0{,}175; p_4 = 0{,}094; p_5 = 0{,}067; p_6 = 0{,}055; p_7 = 0{,}027; p_8 = 0{,}025$ ja kysytty todennäköisyys voidaan laskea multimonijakauman pistetodennäköisyysfunktion arvona

$$\begin{aligned}
&f(2,0,0,1,0,0,0,0) \\
&= P\{X_1 = 2, X_2 = 0, X_3 = 0, X_4 = 1, X_5 = 0, X_6 = 0, X_7 = 0, X_8 = 0\} \\
&= \frac{3!}{2!0!0!1!0!0!0!0!} \cdot 0{,}370^2 \cdot 0{,}188^0 \cdot 0{,}175^0 \cdot 0{,}094^1 \cdot 0{,}067^0 \cdot 0{,}055^0 \cdot 0{,}027^0 \cdot 0{,}025^0 \\
&= 3 \cdot 0{,}370^2 \cdot 0{,}094 \approx \underline{0{,}039}.
\end{aligned}$$

Huomautus 4.4. Esimerkissä 4.3 multinomijakauman parametrit $p_1, \ldots, p_8$ olivat täysin tunnettuja, sillä vaalit olivat jo ohi ja äänien jakautuminen ehdokkaiden kesken tiedettiin. Tilanne, jossa parametri tunnetaan, on kuitenkin harvinainen eikä useinkaan kovin kiinnostava. Usein tilastollisessa päättelyssä onkin tarkoituksena estimoida tuntematonta parametria otoksen perusteella.

4.2 Dirichlet-jakauma multinomijakauman parametrien priorina

Jotta multinomijakauman tuntematonta parametrivektoria voidaan käsitellä satunnaismuuttujana, täytyy sille muodostaa priorijakauma. Dirichlet-jakauma on beta-jakauman moniulotteinen yleistys ja sitä käytetään bayesilaisessa tilastotieteessä priorijakaumana multinomijakauman parametreille [3, s. 83].

Määritelmä 4.5. Olkoon $\boldsymbol{X} = (X_1, \ldots, X_k)$ k-ulotteinen satunnaisvektori, jolle

$$0 < X_i < 1 \quad \text{kaikilla } i \in \{1, \ldots, k\} \quad \text{ja} \quad \sum_{i=1}^{k} X_i = 1.$$

Satunnaisvektori $\boldsymbol{X}$ noudattaa $(k-1)$-ulotteista Dirichlet-jakaumaa parametrilla $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ ($\alpha_i \in \mathbb{R}$ ja $\alpha_i > 0$ kaikilla $i \in \{1, \ldots, k\}$), merkitään $\boldsymbol{X} \sim Dir(\alpha_1, \ldots, \alpha_k)$, jos sen tiheysfunktio on muotoa

$$f(x_1, \ldots, x_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}. \tag{4.6}$$

Olkoon satunnaisvektorin $\boldsymbol{X}$ jakauma $Mult(n; p_1, \ldots, p_k)$, jossa parametrit $p_1, \ldots, p_k$ ovat tuntemattomia. Merkitään multinomikokeen perusjoukkoa $\Omega = (E_1, \ldots, E_k)$. Bayesilaisilla menetelmillä parametrivektoria $\boldsymbol{p}$ voidaan estimoida muodostamalla parametrille

ensin Dirichlet-priorijakauma parametrilla $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$. Priorijakauman parametreja $\alpha_1, \ldots, \alpha_k$ kutsutaan *hyperparametreiksi*, jotta ne erotetaan parametreista $p_1, \ldots, p_k$.

Priorijakauman hyperparametrit $\alpha_1, \ldots, \alpha_k$ ilmaisevat ennakkotietoja parametreista $p_1, \ldots, p_k$, eikä ole yksiselitteistä tapaa valita niitä. Prioriparametrin α_i arvo kertoo tapahtuman E_i suhteellisen uskottavuuden *ennen* havaintoa $\boldsymbol{x}$ [5, s. 784]. Ennakkotiedot voivat perustua aiempiin havaintoihin tai esimerkiksi asiantuntijoiden subjektiiviseen tietämykseen [5, s. 781].

Kun priorijakauma on muodostettu, poimitaan otos $\boldsymbol{x} \mid \boldsymbol{p}$ multinomijakaumasta. Priorijakaumasta ja otoksen multinomisesta uskottavuudesta voidaan muodostaa Bayesin kaavan avulla posteriorijakauma parametrille $\boldsymbol{p} \mid \boldsymbol{x}$. Posteriorijakauman perusteella voidaan lopuksi tehdä johtopäätöksiä parametrista siten, että huomioidaan sekä ennakkotieto että havainto.

Jos prioritietoa ei ole saatavilla tai sitä on vähän, voidaan käyttää epäinformatiivista prioria, jolloin kaikkia alkeistapauksia E_i pidetään ennakkoon yhtä todennäköisinä. Epäinformatiivinen priori voidaan valita myös esimerkiksi silloin, jos halutaan minimoida prioritiedon vaikutus posterioritietoon, jotta yksittäisen havainnon vaikutus johtopäätöksiin voidaan tehdä mahdollisimman suureksi. Epäinformatiiviseksi prioriksi voidaan valita esimerkiksi tasainen *Bayes-Laplace-priori*, jossa $\alpha_i = 1$ kaikilla $i \in \{1, \ldots, k\}$. Sitä käytetään tutkielman seuraavissa esimerkeissä. Toinen esimerkki epäinformatiivisesta priorista on tasainen *Jeffreysin priori*, jossa $\alpha_i = 1/2$ kaikilla $i \in \{1, \ldots, k\}$. [3, s. 61, 63]

Johtopäätökset parametrista perustetaan siis ainoastaan posteriorijakaumaan.

Lause 4.7. *Olkoon $\boldsymbol{X} = (X_1, \ldots, X_k)$ satunnaisvektori ja $\boldsymbol{X} \mid \boldsymbol{p} \sim Mult(n; p_1, \ldots, p_k)$. Jos parametrilla $\boldsymbol{p}$ on Dirichlet-jakauma hyperparametrivektorilla $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, merkitään $\boldsymbol{p} \sim Dir(\alpha_1, \ldots, \alpha_k)$, niin satunnaisvektorille $\boldsymbol{p} \mid \boldsymbol{x}$ pätee*

$$\boldsymbol{p} \mid \boldsymbol{x} \sim Dir(\alpha_1 + x_1, \ldots, \alpha_k + x_k). \tag{4.8}$$

Todistus. Lauseen 3.8 mukaan posteriorijakauma lasketaan

$$f(\boldsymbol{p} \mid \boldsymbol{x}) = \frac{g(\boldsymbol{x} \mid \boldsymbol{p}) h(\boldsymbol{p})}{\int g(\boldsymbol{x} \mid \boldsymbol{p}) h(\boldsymbol{p}) \mathrm{d}\boldsymbol{p}}.$$

Lauseen 4.1 ja Määritelmän 4.5 mukaan

$$g(\boldsymbol{x} \mid \boldsymbol{p}) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \quad \text{ja} \quad h(\boldsymbol{p}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}.$$

Lasketaan nimittäjän integraali $m(\boldsymbol{x}) = \int g(\boldsymbol{x} \mid \boldsymbol{p})h(\boldsymbol{p})\mathrm{d}\boldsymbol{p}$:

$$\begin{aligned}
m(\boldsymbol{x}) &= \int \frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1} \mathrm{d}\boldsymbol{p} \\
&= \frac{n!}{\prod_{i=1}^k x_i!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \int \prod_{i=1}^k p_i^{x_i} \prod_{i=1}^k p_i^{\alpha_i - 1} \mathrm{d}\boldsymbol{p} \\
&= \frac{n!}{\prod_{i=1}^k x_i!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \int \prod_{i=1}^k p_i^{x_i + \alpha_i - 1} \mathrm{d}\boldsymbol{p} \\
&= \frac{n!}{\prod_{i=1}^k x_i!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{\prod_{i=1}^k \Gamma(x_i + \alpha_i)}{\Gamma(\sum_{i=1}^k (x_i + \alpha_i))} \frac{\Gamma(\sum_{i=1}^k (x_i + \alpha_i))}{\prod_{i=1}^k \Gamma(x_i + \alpha_i)} \int \prod_{i=1}^k p_i^{x_i + \alpha_i - 1} \mathrm{d}\boldsymbol{p} \\
&= \frac{n!}{\prod_{i=1}^k x_i!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{\prod_{i=1}^k \Gamma(x_i + \alpha_i)}{\Gamma(\sum_{i=1}^k (x_i + \alpha_i))} \int \frac{\Gamma(\sum_{i=1}^k (x_i + \alpha_i))}{\prod_{i=1}^k \Gamma(x_i + \alpha_i)} \prod_{i=1}^k p_i^{x_i + \alpha_i - 1} \mathrm{d}\boldsymbol{p}.
\end{aligned}$$

Määritelmän 4.5 ja Määritelmän 2.10 kohdan (iii) mukaan integraalille pätee

$$\int \frac{\Gamma(\sum_{i=1}^k (x_i + \alpha_i))}{\prod_{i=1}^k \Gamma(x_i + \alpha_i)} \prod_{i=1}^k p_i^{x_i + \alpha_i - 1} \mathrm{d}\boldsymbol{p} = 1,$$

joten saadaan

$$m(\boldsymbol{x}) = \frac{n!}{\prod_{i=1}^k x_i!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{\prod_{i=1}^k \Gamma(x_i + \alpha_i)}{\Gamma(\sum_{i=1}^k (x_i + \alpha_i))}.$$

Sijoittamalla $g(\boldsymbol{x} \mid \boldsymbol{p})$, $h(\boldsymbol{p})$ ja $m(\boldsymbol{x})$ Lauseessa 3.8 esitettyyn Bayesin kaavaan, saadaan posteriorijakaumaksi lopulta

$$\begin{aligned}
f(\boldsymbol{p} \mid \boldsymbol{x}) &= \frac{g(\boldsymbol{x} \mid \boldsymbol{p})h(\boldsymbol{p} \mid \boldsymbol{\alpha})}{\int g(\boldsymbol{x} \mid \boldsymbol{p})h(\boldsymbol{p} \mid \boldsymbol{\alpha})\mathrm{d}\boldsymbol{p}} = \frac{\left[\frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}\right]}{\left[\frac{n!}{\prod_{i=1}^k x_i!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{\prod_{i=1}^k \Gamma(\alpha_i + x_i)}{\Gamma(\sum_{i=1}^k (\alpha_i + x_i))}\right]} \\
&= \frac{\prod_{i=1}^k p_i^{\alpha_i + x_i - 1}}{\left[\frac{\prod_{i=1}^k \Gamma(\alpha_i + x_i)}{\Gamma(\sum_{i=1}^k (\alpha_i + x_i))}\right]} = \frac{\Gamma(\sum_{i=1}^k (\alpha_i + x_i))}{\prod_{i=1}^k \Gamma(\alpha_i + x_i)} \prod_{i=1}^k p_i^{\alpha_i + x_i - 1}
\end{aligned}$$

ja Määritelmän 4.5 nojalla $\boldsymbol{p} \mid \boldsymbol{x} \sim Dir(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$. □

Dirichlet-jakauma on *konjugaatti priorijakauma* multinomijakauman parametreille. Konjugaattisuudella tarkoitetaan sitä, että jos multinomiparametrien priorijakaumaksi valitaan Dirichlet-jakauma, myös posteriorijakauma on Dirichlet-jakauma [3, s. 40, 582]. Tosin parametrit eroavat priorijakaumaan nähden, kuten huomattiin Lauseesta 4.7.

Kun posteriorijakauma parametrille $\boldsymbol{p} \mid \boldsymbol{x}$ on muodostettu, voidaan johtopäätöksiä tehdä yksittäisestä parametrista p_j laskemalla parametrin p_j reunaposteriorijakauma.

Lause 4.9. *Olkoon k-ulotteinen satunnaisvektori $\boldsymbol{p} = (p_1, \ldots, p_k)$ Dirichlet-jakautunut hyperparametreilla $\alpha_1, \ldots, \alpha_k$, merkitään $\boldsymbol{p} \sim Dir(\alpha_1, \ldots, \alpha_k)$. Tällöin reunajakauma kullekin p_j ($j \in \{1, \ldots, k\}$) on $Beta(\alpha_j, \alpha_0 - \alpha_j)$, jossa $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.*

Todistus. Määritelmän 2.10 kohdan (*iii*) ja Määritelmän 4.5 mukaan vektorin $\boldsymbol{p}$ tiheysfunktiolle pätee

$$\underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{k \text{ kpl}} \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \mathrm{d}p_1 \cdots \mathrm{d}p_k = 1, \tag{4.10}$$

jossa tekijä $\dfrac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)}$ voidaan kirjoittaa muotoon

$$\frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} = \frac{\Gamma(\alpha_0 - \alpha_j)}{\prod_{\substack{i=1 \\ i \neq j}}^{k} \Gamma(\alpha_i)} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)}.$$

Yhtälö (4.10) voidaan siis kirjoittaa muodossa

$$\frac{\Gamma(\alpha_0 - \alpha_j)}{\prod_{\substack{i=1 \\ i \neq j}}^{k} \Gamma(\alpha_i)} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{k-1 \text{ kpl}} \prod_{\substack{i=1 \\ i \neq j}}^{k} p_i^{\alpha_i - 1} \mathrm{d}p_1 \cdots \mathrm{d}p_{j-1} \mathrm{d}p_{j+1} \cdots \mathrm{d}p_k \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)} \int_{-\infty}^{\infty} p_j^{\alpha_j - 1} \mathrm{d}p_j = 1.$$

Jakamalla puolittain tekijällä $\dfrac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)} \int_{-\infty}^{\infty} p_j^{\alpha_j - 1} \mathrm{d}p_j$ saadaan edelleen

$$\frac{\Gamma(\alpha_0 - \alpha_j)}{\prod_{\substack{i=1 \\ i \neq j}}^{k} \Gamma(\alpha_i)} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{k-1 \text{ kpl}} \prod_{\substack{i=1 \\ i \neq j}}^{k} p_i^{\alpha_i - 1} \mathrm{d}p_1 \cdots \mathrm{d}p_{j-1} \mathrm{d}p_{j+1} \cdots \mathrm{d}p_k = \frac{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)}{\Gamma(\alpha_0) \int_{-\infty}^{\infty} p_j^{\alpha_j - 1} \mathrm{d}p_j}. \tag{4.11}$$

Käyttämällä Lauseessa 2.21 annettua Beta-funktion ominaisuutta, sekä vektorin $\boldsymbol{p}$ alkioille Määritelmän 4.5 nojalla pätevää ominaisuutta $1 - p_j = \sum_{\substack{i=1 \\ i \neq j}}^{k} p_i$, voidaan kirjoittaa

$$\begin{aligned}(4.12)\quad \frac{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)}{\Gamma(\alpha_0)} &= \frac{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)}{\Gamma(\alpha_j + \alpha_0 - \alpha_j)} = \int_{-\infty}^{\infty} p_j^{\alpha_j - 1}(1 - p_j)^{\alpha_0 - \alpha_j - 1}\mathrm{d}p_j \\ &= \int_{-\infty}^{\infty} p_j^{\alpha_j - 1}(\sum_{\substack{i=1 \\ i \neq j}}^{k} p_i)^{\alpha_0 - \alpha_j - 1}\mathrm{d}p_j = (\sum_{\substack{i=1 \\ i \neq j}}^{k} p_i)^{\alpha_0 - \alpha_j - 1}\int_{-\infty}^{\infty} p_j^{\alpha_j - 1}\mathrm{d}p_j.\end{aligned}$$

Sijoittamalla yhtälön (4.12) tulos yhtälöön (4.11) saadaan

$$\begin{aligned}(4.13)\quad &\frac{\Gamma(\alpha_0 - \alpha_j)}{\prod_{\substack{i=1 \\ i \neq j}}^{k}\Gamma(\alpha_i)}\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k-1 \text{ kpl}}\prod_{\substack{i=1 \\ i \neq j}}^{k} p_i^{\alpha_i - 1}\mathrm{d}p_1 \cdots \mathrm{d}p_{j-1}\mathrm{d}p_{j+1}\cdots\mathrm{d}p_k \\ &= (\sum_{\substack{i=1 \\ i \neq j}}^{k} p_i)^{\alpha_0 - \alpha_j - 1}\frac{\int_{-\infty}^{\infty} p_j^{\alpha_j - 1}\mathrm{d}p_j}{\int_{-\infty}^{\infty} p_j^{\alpha_j - 1}\mathrm{d}p_j} = (\sum_{\substack{i=1 \\ i \neq j}}^{k} p_i)^{\alpha_0 - \alpha_j - 1} = (1 - p_j)^{\alpha_0 - \alpha_j - 1}.\end{aligned}$$

Määritelmän 2.25 mukaan reunajakauma f_j kullekin p_j saadaan yhteistiheysfunktiosta integroimalla:

$$\begin{aligned}&f_j(p_j \mid p_1, \ldots, p_{j-1}, p_{j+1}, \ldots, p_k) \\ &= \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k-1}\frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k} p_i^{\alpha_i - 1}\mathrm{d}p_1 \cdots \mathrm{d}p_{j-1}\mathrm{d}p_{j+1}\cdots\mathrm{d}p_k \\ &= \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\frac{\prod_{\substack{i=1 \\ i \neq j}}^{k}\Gamma(\alpha_i)}{\Gamma(\alpha_0 - \alpha_j)}p_j^{\alpha_j - 1}\frac{\Gamma(\alpha_0 - \alpha_j)}{\prod_{\substack{i=1 \\ i \neq j}}^{k}\Gamma(\alpha_i)}\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k-1}\prod_{\substack{i=1 \\ i \neq j}}^{k} p_i^{\alpha_i - 1}\mathrm{d}p_1 \cdots \mathrm{d}p_{j-1}\mathrm{d}p_{j+1}\cdots\mathrm{d}p_k \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)}p_j^{\alpha_j - 1}\frac{\Gamma(\alpha_0 - \alpha_j)}{\prod_{\substack{i=1 \\ i \neq j}}^{k}\Gamma(\alpha_i)}\underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{k-1}\prod_{\substack{i=1 \\ i \neq j}}^{k} p_i^{\alpha_i - 1}\mathrm{d}p_1 \cdots \mathrm{d}p_{j-1}\mathrm{d}p_{j+1}\cdots\mathrm{d}p_k.\end{aligned}$$

Kun sijoitetaan vielä yhtälön (4.13) tulos, saadaan reunajakaumaksi f_j kullekin p_j

$$f_j(p_j \mid p_1, \ldots, p_{j-1}, p_{j+1}, \ldots, p_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)\Gamma(\alpha_0 - \alpha_j)}p_j^{\alpha_j - 1}(1 - p_j)^{\alpha_0 - \alpha_j - 1},$$

joka on Määritelmän 2.19 nojalla $Beta(\alpha_j, \alpha_0 - \alpha_j)$. □

Esimerkki 4.14 (Esimerkki bayesilaisen todennäköisyysvälin laskemisesta)**.** MTV:n Research Insight Finland Oy:lta teettämässä haastattelututkimuksessa kysyttiin "Minkä puolueen tai ryhmittymän listalla olevaa ehdokasta äänestäisitte, jos eduskuntavaalit olisivat nyt?". Kysely suoritettiin toukokuussa 2010. Haastateltavia oli 1445, joista 1061 ilmoitti kantansa. Tutkimuksen tulokset on esitetty Taulukossa 4.14. Kannattajamäärät on laskettu annetuista kannatusprosenteista [7]. Mitkä ovat 95%:n todennäköisyysvälit puolueiden todelliselle kannatukselle tutkimuksentekohetkellä?

Taulukko 4.14: Puoluekannatus MTV:n teettämässä tutkimuksessa 21.5.2010

Kok	SDP	Kesk	Vihr	PS	Vas	KD	RKP	muu	yht.
256	219	196	105	92	90	44	40	19	1061
24,1	20,6	18,5	9,9	8,7	8,5	4,1	3,8	1,8	100%

Ratkaisu:

Aineiston keräys voidaan ajatella 1061:n riippumattoman satunnaiskokeen toistona, jossa kussakin valitaan satunnaisesti yksi henkilö ja kysytään, mitä puoluetta hän äänestäisi. Merkitään yksittäisen satunnaiskokeen perusjoukkoa

$$\Omega = \{E_1 = \text{Kok}, E_2 = \text{SDP}, E_3 = \text{Kesk}, E_4 = \text{Vihr}, E_5 = \text{PS}, \\ E_6 = \text{Vas}, E_7 = \text{KD}, E_8 = \text{RKP}, E_9 = \text{muu}\}$$

ja perusjoukon alkeistapauksen E_i esiintymisen todennäköisyyttä $P(E_i) = p_i$. Koska perusjoukon alkeistapaukset ovat erillisiä, satunnaisvektorin $\boldsymbol{p} = (p_1, \ldots, p_9)$ alkioille pätee $\sum_{i=1}^{9} p_i = 1$. Määritellään satunnaisvektori $\boldsymbol{X} = (X_1, \ldots, X_9)$ siten, että X_i on alkeistapauksen E_i esiintymisen lukumäärä, kun koetta toistetaan 1061 kertaa. Selvästi $\sum_{i=1}^{9} X_i = 1061$. Lauseen 4.1 nojalla $\boldsymbol{X} \mid \boldsymbol{p} \sim Mult(1061; p_1, \ldots, p_9)$. Ilmoitetaan kokeesta saatu tilastoaineisto (havainto) vektorina

$$\boldsymbol{x} = (x_1 = 256, x_2 = 219, x_3 = 196, x_4 = 105, x_5 = 92, x_6 = 90, x_7 = 44, x_8 = 40, x_9 = 19).$$

Estimoidaan parametrivektorin $\boldsymbol{p} = (p_1, \ldots, p_9)$ alkioita havainnon perusteella. Oletetaan vektorille $\boldsymbol{p}$ tasainen Dirichlet-priorijakauma. Käytetään Bayes-Laplace-prioria ja merkitään $\boldsymbol{p} \sim Dir(1, 1, 1, 1, 1, 1, 1, 1, 1)$. Hyperparametrivektori on siis

$$\boldsymbol{\alpha} = (\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1, \alpha_4 = 1, \alpha_5 = 1, \alpha_6 = 1, \alpha_7 = 1, \alpha_8 = 1, \alpha_9 = 1).$$

Lauseen 4.7 nojalla parametrivektorin $\boldsymbol{p}$ havainnosta $\boldsymbol{x}$ ehdollinen posteriorijakauma on $Dir(\alpha_1 + x_1, \ldots, \alpha_9 + x_9)$, jolloin Lauseen 4.9 nojalla reunaposteriorijakauma kullekin $p_i \mid \boldsymbol{x}$ on $Beta(\alpha_i + x_i, \sum_{j=1}^{9}(\alpha_j + x_j) - (\alpha_i + x_i))$. Parametreille $\boldsymbol{p} \mid \boldsymbol{x}$ ja $p_i \mid \boldsymbol{x}$ pätee siis

$$\boldsymbol{p} \mid \boldsymbol{x} \sim Dir(257, 220, 197, 106, 93, 91, 45, 41, 20)$$

ja

$$p_1 \mid \boldsymbol{x} \sim Beta(257, 813), \quad p_2 \mid \boldsymbol{x} \sim Beta(220, 850), \quad p_3 \mid \boldsymbol{x} \sim Beta(197, 873),$$
$$p_4 \mid \boldsymbol{x} \sim Beta(106, 964), \quad p_5 \mid \boldsymbol{x} \sim Beta(93, 977), \quad p_6 \mid \boldsymbol{x} \sim Beta(91, 979),$$
$$p_7 \mid \boldsymbol{x} \sim Beta(45, 1025), \quad p_8 \mid \boldsymbol{x} \sim Beta(41, 1029).$$

Parametria p_9 ei tarkastella, sillä se ei ole tehtävänannossa kiinnostuksen kohteena.

Tarkastellaan aluksi parametria p_1. Tällöin 95%:n todennäköisyysväli on yhtälön (3.14) mukaisesti väli $[c, d]$, jonka päätepisteet toteuttavat yhtälön $P\{c < p_1 < d\} = 0{,}95$. Kun parametrin p_1 tiheysfunktio f_1 on Määritelmän 2.19 mukaan

$$f_1(p_1) = \frac{\Gamma(1070)}{\Gamma(257)\Gamma(813)} p_1^{256}(1 - p_1)^{812}, \tag{4.15}$$

niin yhtälön (2.12) mukaan 95% todennäköisyysvälin päätepisteet toteuttavat yhtälön

$$P\{c < p_1 < d\} = \int_c^d \frac{\Gamma(1070)}{\Gamma(257)\Gamma(813)} p_1^{256}(1 - p_1)^{812} \mathrm{d}p_1 = 0{,}95. \tag{4.16}$$

Koska funktio on jatkuva, niin yhtälön toteuttavia lukuja c ja d on ääretön määrä. Lasketaan tässä todennäköisyysväli siten, että päätepisteen c vasemmalle puolelle jää yhtä suuri todennäköisyysmassa kuin päätepisteen d oikealle puolelle. Toisin sanottuna valitaan symmetrinen väli $[c, d]$ siten, että päätepisteille pätee $P\{p_1 < c\} = P\{p_1 > d\} = 0{,}025$. (Huomautetaan, että tällöin yhtälö (4.16) toteutuu, sillä $P\{c < p_1 < d\} = 1 - P(\{p_1 < c\} \cap \{p_1 > d\}) = 1 - (P\{p_1 < c\} + P\{p_1 > d\}) = 1 - (0{,}025 + 0{,}025) = 1 - 0{,}05 = 0{,}95$.)

Yhtälöiden (2.16) ja (2.18) nojalla symmetrisen todennäköisyysvälin päätepisteet toteuttavat siis yhtälöt

$$P\{p_1 < c\} = \int_{-\infty}^{c} \frac{\Gamma(1070)}{\Gamma(257)\Gamma(813)} p_1^{256}(1 - p_1)^{812} \mathrm{d}p_1 = 0{,}025 \quad \text{ja}$$
$$P\{p_1 < d\} = \int_{-\infty}^{d} \frac{\Gamma(1070)}{\Gamma(257)\Gamma(813)} p_1^{256}(1 - p_1)^{812} \mathrm{d}p_1 = 1 - P\{p_1 > d\} = 1 - 0{,}025 = 0{,}975.$$

Käytetään BetaBuster-ohjelmaa yhtälöiden ratkaisemisessa [8]. Ohjelma antaa
$P\{p_1 < c\} = 0{,}025$, kun $c = 0{,}215070233$ ja
$P\{p_1 < d\} = 0{,}975$, kun $d = 0{,}266223490$. Saadaan siis todennäköisyysväliksi $[c, d] = [0{,}215; 0{,}266]$. Siis todennäköisyys, että Kokoomuksen todellinen kannatus on tutkimuksentekohetkellä välillä [0,215;0,266] on 0,95.

Toistetaan sama tarkastelu parametreille $p_2, \ldots, p_8$ ja käytetään jokaisen kohdalla BetaBuster-ohjelmaa todennäköisyysvälien ratkaisemiseksi. Kaikkien parametrien posteriorivälit on annettu Taulukossa 4.16.

Taulukko 4.16: Puoluekannatusten todennäköisyysvälit

Parametri i	$c_i, P\{p_i < c_i\} = 0{,}025$	$d_i, P\{p_i < d_i\} = 0{,}975$	Todennäköisyysväli $[c_i, d_i]$
p_1	0,215070233	0,266223490	[0,215;0,266]
p_2	0,181933403	0,230323836	[0,182;0,230]
p_3	0,161469087	0,207873553	[0,161;0,208]
p_4	0,081896894	0,117652953	[0,082;0,118]
p_5	0,070790321	0,104503319	[0,071;0,105]
p_6	0,069089979	0,102471933	[0,069;0,102]
p_7	0,030867560	0,054863859	[0,031;0,055]
p_8	0,027661217	0,050606526	[0,028;0,051]

4.3 Dirichlet-multinomijakauma

Dirichlet-multinomijakauma on Dirichlet-jakauman ja multinomijakauman yhteisjakauma. Jos satunnaismuuttuja X noudattaa multinomijakaumaa parametreilla n ja $\boldsymbol{p}$ ja parametrilla $\boldsymbol{p}$ on Dirichlet-priorijakauma hyperparametrilla $\boldsymbol{\alpha}$, niin Dirichlet-multinomijakauma on satunnaismuuttujasta X poimitun havainnon $\boldsymbol{x}$ hyperparametristä $\boldsymbol{\alpha}$ ehdollinen jakauma. Jakaumaa käyttäen voidaan siten laskea havainnon todennäköisyys, kun prioritieto multinomijakauman parametreista on annettu.

Lause 4.17. *Olkoon satunnaiskokeella k erillistä tulosvaihtoehtoa, merkitään perusjoukkoa $\Omega = \{1, \ldots, k\}$. Toistaan koetta n kertaa riippumattomasti. Olkoon $\boldsymbol{X} = (X_1, \ldots, X_k)$ k-ulotteinen satunnaisvektori, jossa X_i on niiden toistojen lukumäärä, joissa tulosvaihtoehto $i \in \{1, \ldots, k\}$ on havaittu ja olkoon p_i i:nnen tulosvaihtoehdon todennäköisyys. Siis $\boldsymbol{X} \mid \boldsymbol{p} \sim Mult(n; p_1, \ldots, p_k)$. Oletetaan vektorille $\boldsymbol{p} = (p_1, \ldots, p_k)$ Dirichlet-priorijakauma hyperparametrilla $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_k)$. Nyt satunnaismuuttujalla $\boldsymbol{X}$ on jakauma, jonka pistetodennäköisyysfunktio on muotoa*

$$f(\boldsymbol{x}) = \frac{n!}{\prod_{i=1}^{k} x_i!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Gamma(n + \sum_{i=1}^{k} \alpha_i)} \prod_{i=1}^{k} \frac{\Gamma(x_i + \alpha_i)}{\Gamma(\alpha_i)}. \tag{4.18}$$

Jakaumaa kutsutaan Dirichlet-multinomijakaumaksi. Sanotaan, että $\boldsymbol{X}$ noudattaa $(k-1)$-ulotteista Dirichlet-multinomijakaumaa parametrein n ja $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_k)$, merkitään $\boldsymbol{X} \sim Dirmult(n; \alpha_1, \ldots \alpha_k)$.

Todistus. Lauseen 4.1 nojalla $\boldsymbol{X} \mid \boldsymbol{p} \sim Mult(n; p_1, \ldots, p_k)$ ja satunnaisvektorin $\boldsymbol{X} \mid \boldsymbol{p}$ pistetodennäköisyysfunktio on muotoa

$$g(\boldsymbol{x} \mid \boldsymbol{p}) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i}.$$

Koska $\boldsymbol{p} \sim Dir(\alpha_1, \ldots, \alpha_k)$, niin Määritelmän 4.5 nojalla satunnaismuuttujan $\boldsymbol{p}$ tiheysfunktio on muotoa

$$h(\boldsymbol{p}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}. \tag{4.19}$$

Kun g on satunnaisvektorin $\boldsymbol{X} \mid \boldsymbol{p}$ uskottavuusfunktio ja h satunnaisvektorin $\boldsymbol{p}$ prioritodennäköisyysfunktio, niin satunnaisvektorin $\boldsymbol{X}$ hyperparametrista $\boldsymbol{\alpha}$ ehdollisen jakauman tiheysfunktio on Lauseen 3.8 nojalla reunauskottavuusfunktio, joka saadaan integroimalla tulofunktio

$$g(\boldsymbol{x} \mid \boldsymbol{p})h(\boldsymbol{p})$$

parametrin $\boldsymbol{p}$ yli:

$$\begin{aligned}\int g(\boldsymbol{x} \mid \boldsymbol{p})h(\boldsymbol{p})\mathrm{d}\boldsymbol{p} &= \int \frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}\mathrm{d}\boldsymbol{p} \\ &= \frac{n!}{\prod_{i=1}^k x_i!} \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{\prod_{i=1}^k \Gamma(\alpha_i + x_i)}{\Gamma(\sum_{i=1}^k (\alpha_i + x_i))}.\end{aligned}$$

(Integraali on laskettu aiemmin Lauseen 4.7 todistuksessa.) □

Dirichlet-multinomijakaumaa voidaan soveltaa esimerkiksi bayesilaisessa mallivertailussa.

Esimerkki 4.20 (Esimerkki bayesilaisesta mallivertailusta)**.** Vuosina 2006-2012 peruskoulun yhdeksännen luokan päätti Suomessa 448418 nuorta. Heistä poikia oli 228397 ja tyttöjä 220021. 209191 poikaa ja 200855 tyttöä jatkoi heti opintojaan joko lukiossa tai toisen asteen ammatillisessa koulutuksessa. Sen sijaan 19206 poikaa ja 19166 tyttöä ei heti jatkanut tutkintotavoitteista opiskelua, eli lopetti opinnot tai aloitti opinnot peruskoulun 10. luokalla. [9]

Aineisto on taulukoitu sukupuolen ja tutkintotavoitteiseen koulutukseen sijoittumisen mukaan Taulukossa 4.20. Onko sukupuolella ja tutkintotavoitteisiin opintoihin sijoittumisella riippuvuutta aineiston perusteella?

Taulukko 4.20: Peruskoulun yhdeksännen luokan päättäneiden sijoittuminen jatko-opintoihin vuosina 2006-2012

Sukupuoli	Lukio/ammatillinen koulutus	Ei jatka/10. luokka	yht.
Poika	209191	19206	228397
Tyttö	200855	19166	220021

Ratkaisu:

Merkitään satunnaismuuttujia $S =$ "sukupuoli" ja $T =$ "jatko-opinnot".

Käytetään kahta eri mallia tilastoaineiston todennäköisyyden laskemiseksi. Määritelään mallit:

$$M_1 : \text{Satunnaismuuttujat } S \text{ ja } T \text{ ovat riippuvia ja}$$
$$M_2 : \text{Satunnaismuuttujat } S \text{ ja } T \text{ ovat riippumattomia.}$$

Lasketaan aluksi tilastoaineiston todennäköisyys mallilla M_1.

Määritellään kaksiulotteinen satunnaismuuttuja (S, T) siten, että

$$S = \begin{cases} 1, & \text{jos henkilö on poika,} \\ 2, & \text{jos henkilö on tyttö,} \end{cases}$$

$$T = \begin{cases} 1, & \text{jos henkilö jatkaa tutkintotavoitteista opiskelua heti,} \\ 2, & \text{jos henkilö ei jatka tutkintotavoitteista opiskelua heti.} \end{cases}$$

Satunnaismuuttujan perusjoukko on $\Omega_{S,T} = \{(1,1), (1,2), (2,1), (2,2)\}$.

Merkitään perusjoukon alkeistapauksiin liittyviä todennäköisyyksiä

$$P\{S = 1, T = 1\} = p_{11}, \quad P\{S = 1, T = 2\} = p_{12},$$
$$P\{S = 2, T = 1\} = p_{21}, \quad P\{S = 2, T = 2\} = p_{22}.$$

Selvästi $p_{11} + p_{12} + p_{21} + p_{22} = 1$ ja $p_{ij} > 0$ kaikilla $i, j \in \{1, 2\}$.

Määritellään satunnaisvektori $\boldsymbol{X} = (X_{11}, X_{12}, X_{21}, X_{22})$ siten, että X_{ij} on alkeistapauksen (i, j) esiintymisen lukumäärä. Alkioille X_{ij} pätee $X_{11} + X_{12} + X_{21} + X_{22} = n$, jossa $n = 448418$ on kaikkien vuosina 2006-2012 peruskoulun yhdeksännen luokan päättäneiden lukumäärä. Lauseen 4.1 nojalla $\boldsymbol{X} \mid \boldsymbol{p} \sim Mult(448418; p_{11}, p_{12}, p_{21}, p_{22})$. Asetetaan parametrivektorille $\boldsymbol{p}$ tasainen Bayes-Laplace-priorijakauma. Hyperparametrivektori on siten $\boldsymbol{\alpha} = (\alpha_{11} = 1, \alpha_{12} = 1, \alpha_{21} = 1, \alpha_{22} = 1)$, merkitään $\boldsymbol{p} \sim Dir(1, 1, 1, 1)$. Lauseen 4.17 nojalla satunnaismuuttuja $\boldsymbol{X}$ noudattaa Dirichlet-multinomijakaumaa, merkitään $\boldsymbol{X} \sim Dirmult(448418; 1, 1, 1, 1)$. Riippuvuuden sallivalla mallilla M_1 havainnon $\boldsymbol{x} = (209191, 19206, 200855, 19166)$ todennäköisyys voidaan siten Määritelmän 2.1 ja Lauseen 4.17 mukaan laskea:

$$
\begin{aligned}
P(\boldsymbol{x} \mid M_1) &= P\{X_{11} = 209191, X_{12} = 19206, X_{21} = 200855, X_{22} = 19166\} \\
&= \frac{448418!}{209191! \cdot 19206! \cdot 200855! \cdot 19166!} \frac{\Gamma(4)}{\Gamma(448422)} \frac{\Gamma(209192)\Gamma(19207)\Gamma(200856)\Gamma(19167)}{\Gamma(1)\Gamma(1)\Gamma(1)\Gamma(1)} \\
&= \frac{448418! \cdot 3! \cdot 209191! \cdot 19206! \cdot 200855! \cdot 19166!}{209191! \cdot 19206! \cdot 200855! \cdot 19166! \cdot 448421 \cdot 448420 \cdot 448419 \cdot 448418!} \\
&= \frac{6}{448421 \cdot 448420 \cdot 448419}.
\end{aligned}
$$

Lasketaan seuraavaksi tilastoaineiston todennäköisyys mallilla M_2.

Jos satunnaismuuttujat S ja T ovat riippumattomia, voidaan niiden yhteispistetodennäköisyysfunktio $P(S,T)$ Määritelmän 2.28 nojalla laskea reunajakaumien pistetodennäköisyysfunktioiden tulona $P(S)P(T)$. Määritellään reunajakaumat.

Merkitään satunnaismuuttujien S ja T perusjoukkoja

$$
\begin{aligned}
\Omega_S &= \{1, 2 \mid 1 = \text{mies}, 2 = \text{nainen}\}, \\
\Omega_T &= \{1, 2 \mid 1 = \text{Lukio/ammatillinen koulutus}, 2 = \text{Ei jatka/10. luokka}\}
\end{aligned}
$$

ja perusjoukkojen alkeistapauksin liittyviä todennäköisyyksiä

$$
\begin{aligned}
&P\{S = 1\} = p_1, \quad P\{S = 2\} = p_2, \\
&P\{T = 1\} = q_1 \quad \text{ja} \quad P\{T = 2\} = q_2.
\end{aligned}
$$

Selvästi $p_1 + p_2 = q_1 + q_2 = 1$, ja $p_i, q_i > 0$ kaikilla $i \in \{1, 2\}$.

Määritellään satunnaisvektori $\boldsymbol{Y} = (Y_1, Y_2)$ siten, että Y_i on alkeistapauksen $S = i$ esiintymisen lukumäärä. Alkioille pätee $Y_1 + Y_2 = n$, jossa $n = 448418$ on kaikkien kyseisinä vuosina peruskoulun yhdeksännen luokan päättäneiden lukumäärä. Lauseen 4.1 nojalla $\boldsymbol{Y} \mid \boldsymbol{p} \sim Mult(448418; p_1, p_2)$. Mallien priorioletuksien tulee olla mallivertailussa yhtenevät. Koska parametrivektorille $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ asetettiin mallissa M_1 Dirichlet-priorijakauma hyperparametrivektorilla $\boldsymbol{\alpha} = (\alpha_{11} = 1, \alpha_{12} = 1, \alpha_{21} = 1, \alpha_{22} = 1)$ ja $p_1 = p_{11} + p_{12}$ ja $p_2 = p_{21} + p_{22}$, niin parametrille $\boldsymbol{p}_S = (p_1, p_2)$ asetettavan Dirichlet-priorijakauman hyperparametrivektori on $\boldsymbol{\alpha}_S = (\alpha_{11} + \alpha_{12}, \alpha_{21} + \alpha_{22})$. Merkitään $\boldsymbol{p}_S \sim Dir(2, 2)$. Lauseen 4.17 nojalla satunnaisvektori $\boldsymbol{Y}$ noudattaa Dirichlet-multinomijakaumaa, merkitään $\boldsymbol{Y} \sim Dirmult(448418; 2, 2)$. Nyt voidaan laskea tilastoai-

neiston $\boldsymbol{y} = (228397, 220021)$ todennäköisyys Määritelmän 2.1 ja Lauseen 4.17 mukaan

$$\begin{aligned} P(\boldsymbol{y}) &= P\{Y_1 = 228397, Y_2 = 220021\} \\ &= \frac{448418!}{228397! \cdot 220021!} \frac{\Gamma(4)}{\Gamma(448422)} \frac{\Gamma(228399)\Gamma(220023)}{\Gamma(2)(2)} \\ &= \frac{448418! \cdot 3! \cdot 228398 \cdot 228397! \cdot 220022 \cdot 220021!}{228397! \cdot 220021! \cdot 448421 \cdot 448420 \cdot 448419 \cdot 448418!} \\ &= \frac{6 \cdot 228398 \cdot 220022}{448421 \cdot 448420 \cdot 448419}. \end{aligned}$$

Määritellään satunnaisvektori $\boldsymbol{Z} = (Z_{11}, Z_{12}, Z_{21}, Z_{22})$ siten, että Z_{ij} on ehdollisen tapahtuman $T = i \mid S = j$ esiintymisen lukumäärä. Nähdään, että $Z_{11} + Z_{21} = k$, jossa $k = 228397$ on kaikkien kyseisinä vuosina peruskoulun yhdeksännen luokan päättäneiden poikien lukumäärä ja vastaavasti $Z_{12} + Z_{22} = h$, jossa $h = 220021$, on peruskoulun yhdeksännen luokan päättäneiden tyttöjen lukumäärä. Lauseen 4.1 nojalla satunnaisvektorin $\boldsymbol{Z}$ osajoukoille pätee $(Z_{11}, Z_{21}) \mid \boldsymbol{g}_T \sim Mult(228397; q_1, q_2)$ ja $(Z_{12}, Z_{22}) \mid \boldsymbol{g}_T \sim Mult(220021; q_1, q_2)$. (Huomautetaan, että ehdollisuus satunnaismuuttujasta S vaikuttaa ainoastaan multinomikertoimeen, ei todennäköisyysvektoriin $\boldsymbol{q}_T = (q_1, q_2)$.) Koska osajoukot ovat erillisiä, voidaan satunnaisvektorin $\boldsymbol{Z} \mid \boldsymbol{q}_T$ todennäköisyysjakauma määrittää osajoukkojen jakaumien tulona ja saadaan havainnon $\boldsymbol{z} = (209191, 200855, 19206, 19166)$ parametrista $\boldsymbol{g}_T$ ehdolliseksi todennäköisyydeksi

$$\begin{aligned} P(\boldsymbol{z} \mid \boldsymbol{q}_T) &= P\{Z_{11} = 209191, Z_{12} = 200855, Z_{21} = 19206, Z_{22} = 19166 \mid \boldsymbol{q}_T\} \\ &= P\{Z_{11} = 209191, Z_{21} = 19206 \mid \boldsymbol{q}_T\} P\{Z_{12} = 200855, Z_{22} = 19166 \mid \boldsymbol{q}_T\} \\ &= \frac{228397!}{209191! \cdot 19206!} q_1^{209191} q_2^{19206} \frac{220021!}{200855! \cdot 19166!} q_1^{200855} q_2^{19166} \\ &= \frac{228397! \cdot 220021!}{209191! \cdot 19206! \cdot 200855! \cdot 19166!} q_1^{410046} q_2^{38372}. \end{aligned}$$

Oletetaan parametrille $\boldsymbol{q}_T$ Dirichlet-priorijakauma. Koska $q_1 = p_{11} + p_{21}$ ja $q_2 = p_{12} + p_{22}$ ja $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22}) \sim Dir(1, 1, 1, 1)$, niin parametrivektori $\boldsymbol{g}_T \sim Dir(2, 2)$, jossa $\boldsymbol{\alpha}_T = (\alpha_{11} + \alpha_{21}, \alpha_{12} + \alpha_{22}) = (2, 2)$ on hyperparametrivektori. Havainnon $\boldsymbol{z} = (209191, 200855, 19206, 19166)$ hyperparametrista $\boldsymbol{\alpha}_T$ ehdollinen todennäköisyys saadaan integroimalla uskottavuusfunktion ja prioritodennäköisyysfunktion tulo parametrivektorin $\boldsymbol{g}_T$ yli. Saadaan

$$\begin{aligned} &P(\boldsymbol{z}) = P\{Z_{11} = 209191, Z_{12} = 200855, Z_{21} = 19206, Z_{22} = 19166\} \\ &= \int \frac{228397! \cdot 220021!}{209191! \cdot 19206! \cdot 200855! \cdot 19166!} q_1^{410046} q_2^{38372} \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} q_1 q_2 \mathrm{d}\boldsymbol{q}_T \\ &= \frac{228397! \cdot 220021! \cdot \Gamma(4)\Gamma(410048)\Gamma(38374)}{209191! \cdot 19206! \cdot 200855! \cdot 19166! \cdot \Gamma(2)\Gamma(2)\Gamma(448422)} \int \frac{\Gamma(448422)}{\Gamma(410048)\Gamma(38374)} q_1^{410047} q_2^{38373} \mathrm{d}\boldsymbol{q}_T. \end{aligned}$$

Määritelmän 2.10 kohdan (*iii*) ja Määritelmän 4.5 nojalla integraalille pätee

$$\int \frac{\Gamma(448422)}{\Gamma(410048)\Gamma(38374)} q_1^{410047} q_2^{38373} \mathrm{d}\boldsymbol{q}_T = 1$$

ja saadaan

$$\begin{aligned}
P(\boldsymbol{z}) &= \frac{228397! \cdot 220021! \cdot \Gamma(4)\Gamma(410048)\Gamma(38374)}{209191! \cdot 19206! \cdot 200855! \cdot 19166! \cdot \Gamma(2)\Gamma(2)\Gamma(448422)} \\
&= \frac{228397! \cdot 220021! \cdot 3! \cdot \Gamma(410048)\Gamma(38374)}{209191! \cdot 19206! \cdot 200855! \cdot 19166! \cdot \Gamma(448422)} \\
&= \frac{228397! \cdot 220021! \cdot 6 \cdot 410047! \cdot 38373!}{209191! \cdot 19206! \cdot 200855! \cdot 19166! \cdot 448421 \cdot 448420!} \\
&= \frac{6}{448421} \binom{228397}{19206} \binom{220021}{19166} \left[\binom{448420}{38373} \right]^{-1} \\
&= \frac{6 \cdot \binom{228397}{19206} \binom{220021}{19166}}{448421 \cdot \binom{448420}{38373}}.
\end{aligned}$$

Mallilla M_2, eli riippumattomuusoletuksen alla, saadaan havainnon $\boldsymbol{x} = (209191, 19206, 200855, 19166)$ todennäköisyydeksi:

$$\begin{aligned}
P(\boldsymbol{x} \mid M_2) = P(\boldsymbol{y})P(\boldsymbol{z}) &= \frac{6 \cdot 228398 \cdot 220022}{448421 \cdot 448420 \cdot 448419} \cdot \frac{6 \cdot \binom{228397}{19206} \binom{220021}{19166}}{448421 \cdot \binom{448420}{38373}} \\
&= \frac{36 \cdot 228398 \cdot 220022 \cdot \binom{228397}{19206} \binom{220021}{19166}}{448421^2 \cdot 448420 \cdot 448419 \cdot \binom{448420}{38373}}.
\end{aligned}$$

Mallien prioritodennäköisyydet ovat $P(M_1) = P(M_2) = 0{,}5$, eli kumpaakaan mallia ei suosita etukäteen. Lasketaan posterioritodennäköisyys mallille M_1 yhtälön (3.15) mukaisesti

$$\begin{aligned}
P(M_1 \mid \boldsymbol{x}) &= \frac{P(\boldsymbol{x} \mid M_1)P(M_1)}{P(\boldsymbol{x} \mid M_1)P(M_1) + P(\boldsymbol{x} \mid M_2)P(M_2)} \\
&= \frac{\left[\dfrac{6}{448421 \cdot 448420 \cdot 448419} \cdot 0{,}5 \right]}{\left[\dfrac{6}{448421 \cdot 448420 \cdot 448419} \cdot 0{,}5 \right] + \left[\dfrac{36 \cdot 228398 \cdot 220022 \cdot \binom{228397}{19206} \binom{220021}{19166}}{448421^2 \cdot 448420 \cdot 448419 \cdot \binom{448420}{38373}} \cdot 0{,}5 \right]} \\
&\approx \underline{0{,}75312}.
\end{aligned}$$

Mallin M_2 posterioritodennäköisyys on siten $P(M_2 \mid \boldsymbol{x}) = 1 - 0{,}75312 = 0{,}24688$. Jotta saadaan sanallinen tulkinta todennäköisyyksille, lasketaan Bayes-tekijä. Lasketaan se mallin M_1 puolesta:

$$\begin{aligned}
\text{Bayes-tekijä}(M_1; M_2) &= \frac{P(\boldsymbol{x} \mid M_1)}{P(\boldsymbol{x} \mid M_2)} \\
&= \frac{\left[\dfrac{6}{448421 \cdot 448420 \cdot 448419}\right]}{\left[\dfrac{36 \cdot 228398 \cdot 220022 \cdot \binom{228397}{19206}\binom{220021}{19166}}{448421^2 \cdot 448420 \cdot 448419 \cdot \binom{448420}{38373}}\right]} \\
&= \frac{448421 \cdot \binom{448420}{38373}}{6 \cdot 228398 \cdot 220022 \cdot \binom{228397}{19206}\binom{220021}{19166}} \\
&\approx \underline{3{,}0506}.
\end{aligned}$$

Koska Bayes-tekijä$(M_1; M_2) \in [3, 20]$, niin Taulukon 3.5 perusteella voidaan sanoa, että näyttö mallia M_2 vastaan on todellinen. Tilastoaineisto antaa siis tukea enemmän mallille M_1 ja satunnaismuuttujilla on riippuvuutta tilaston perusteella. Riippuvuuden sallivan mallin posterioritodennäköisyydeksi saadaan $P(M_1 \mid \boldsymbol{x}) = 0{,}75312$.

Vertailun vuoksi sanottakoon, että frekventistisellä χ^2-riippumattomuustestillä χ^2-testisuureen arvoksi saadaan $\chi^2 \approx 13{,}0570$ ja nollahypoteesi *"Satunnaismuuttujat ovat riippumattomia"* hylätään merkitsevyystasolla 0,001. (1-vapausasteisen χ^2-jakauman kriittinen arvo merkitsevyystasolla 0,001 on 10,8274.)

Lähdeluettelo

[1] P. Tuominen: *Todennäköisyyslaskenta I,* 8. painos, Limes ry, Helsinki, 2007.

[2] M.H. Gegroot, M.J. Schervish: *Probability and Statistics,* 3. painos, Addison-Wesley, Boston, 2002.

[3] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin: *Bayesian Data Analysis,* 2. painos, Chapman Hall/CRC, Boca Raton, Florida, 2004.

[4] D. Heckerman: *A Tutorial on Learning With Bayesian Networks,* Technical Report MSR-TR-96-06, Microsoft Research, Advanced Technology Division, 1995.
http://research.microsoft.com/pubs/69588/tr-95-06.pdf

[5] R.E. Kass, A.E. Raftery: *Bayes Factors,* Journal of the American Statistical Association, vol. 90, nro. 430, 773-795, 1995.
http://www.stat.cmu.edu/ kass/papers/bayesfactors.pdf

[6] Suomen virallinen tilasto (SVT): *Presidentinvaalit 2012, I vaali*, Tilastokeskus, Helsinki.
http://www.stat.fi/til/pvaa/2012/01/pvaa_2012_01_2012-01-26_tie_001_fi.html

[7] MTV3: *MTV:n kysely: Keskustan kannatus laskenut,* 29.11.2010.
http://www.mtv.fi/uutiset/kotimaa/artikkeli/mtv-n-kysely–keskustan-kannatus-laskenut–/1852312

[8] S. Chun-Lung: *Betabuster-ohjelmisto.*
http://www.epi.ucdavis.edu/diagnostictests/betabuster.html

[9] Suomen virallinen tilasto (SVT): Koulutukseen hakeutuminen 2006-2012, Tilastokeskus, Helsinki.